Kodikologie und Paläographie im digitalen Zeitalter

Codicology and Palaeography in the Digital Age

Schriften des
Instituts für Dokumentologie und Editorik

Band 2

Schriften des Instituts für Dokumentologie und Editorik — Band 2

Kodikologie und Paläographie im digitalen Zeitalter

Codicology and Palaeography in the Digital Age

herausgegeben von | edited by

Malte Rehbein, Patrick Sahle, Torsten Schaßan

unter Mitarbeit von | in collaboration with

Bernhard Assmann, Franz Fischer, Christiane Fritze

2009
BoD, Norderstedt

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über **http://dnb.d-nb.de/** abrufbar.

Leicht veränderte Fassung für die digitale Publikation (siehe Vorwort).

Slightly modified version to be published digitally (see preface).

Publication réalisée avec le soutien d'Apices
Association Paléographique Internationale
Culture – Écriture – Société
Dotation J.M.M. Hermans.
http://www.palaeographia.org/apices/

Herstellung und Verlag: Books on Demand GmbH, Norderstedt
ISBN: 978-3-8370-9842-6
Einbandgestaltung: Katharina Weber
Satz: XƎTEX und Bernhard Assmann

Inhaltsverzeichnis – Contents

Kodikologie: Vom Katalog zur Forschungsumgebung
Codicology: From Catalogue to Virtual Research Environment

Paläographie: Vom virtuellen Lernen zu neuen Perspektiven
Palaeography: From eLearning to New Research Horizons

Appendizes
Appendices

Vorwort

Die technologischen Entwicklungen seit dem Ende des 20. Jahrhunderts lassen vom Beginn eines »digitalen Zeitalters« sprechen. Wie für andere wissenschaftliche Disziplinen stellt sich auch für die Arbeit mit Handschriften die Frage, welche Bedeutung der Übergang zum Digitalen hat, welchen Einfluss das Internet und der allgegenwärtige Einsatz von Computertechnologie auf die Forschung ausübt und welche Möglichkeiten die schier grenzenlosen Speicher- und Rechenkapazitäten ebenso wie fortgeschrittene Scan- und Bildverarbeitungsverfahren oder das Internet als neue zentrale Publikationsplattform bieten. Wo bestehen Chancen, wo vielleicht sogar Risiken?

Bereits heute werden die Errungenschaften des »digitalen Zeitalters« vielseitig genutzt: Informationsressourcen wie Handschriftenkataloge und Wasserzeichenverzeichnisse werden digitalisiert und in übergreifenden Portalen vereint. Die Virtualisierung des Zugangs erlaubt neue thematische Zusammenstellungen und die Rekonstruktion historischer Sammlungen. Die vollständige digitale Faksimilierung von Handschriften erlaubt einen ungleich leichteren Zugriff auf die Dokumente und führt zu neuen Nutzungsweisen. Die Möglichkeiten des eLearning erweitern die didaktischen Optionen bei der Vermittlung paläographischer Lesekompetenz. Eine schon früher angestrebte, messende, quantifizierende Paläographie kann jetzt endlich auf des Basis eines weit verfügbaren Bildmaterials und mit Hilfe des Computers algorithmisch realisiert werden. Buchstabenmodelle und schreiberspezifische Schriften können erst jetzt genau bestimmt und modelliert werden. Dies eröffnet Perspektiven für eine computergestützte Schrifterkennung und damit neue Grundlagen für Transkriptionen und Editionen, die in der Publikation nicht mehr von den Handschriften abgeschnitten sind, sondern unmittelbar an sie zurück gebunden bleiben.

Wie geht es weiter? Wie weit entfernt sind wir von einer automatischen Texterkennung mittelalterlicher Manuskripte? Wird es in absehbarer Zukunft einen allumfassenden virtuellen Katalog geben, der den Zugriff auf ein vollständig digitalisiertes europäisches oder gar weltweites Handschriftenerbe über das Internet ermöglicht? Entstehen auf dieser Basis kollaborative Plattformen, in denen die Überlieferung erschlossen, transkribiert und ediert wird? Werden sich unter diesen Voraussetzungen die Forschungsmethoden grundlegend ändern und zu neuen Erkenntnissen über die Entwicklung der menschlichen Kultur führen?

Die Verwirklichung alter Träume auf der einen und die Entdeckung neuer Fragehorizonte auf der anderen Seite kennzeichnen eine Situation des raschen Fortschritts und des tiefgreifenden Wandels in der kodikologischen und paläographischen Theorie und Praxis. Um den Stand der Forschung zu dokumentieren und Entwicklungsperspektiven aufzuzeigen, hat das Institut für Dokumentologie und Editorik (IDE) – ein

Zusammenschluss junger Wissenschaftlerinnen und Wissenschaftler aus dem Bereich des Humanities Computing – Ende Oktober 2008 einen Call for Papers für den vorliegenden Sammelband veröffentlicht. Auf der Basis von insgesamt 31 eingereichten Vorschlägen werden hier nun 21 Beiträge veröffentlicht, welche Arbeiten und Projekte von Forschern aus Australien, Deutschland, Frankreich, Großbritannien, Irland, Italien, Österreich, Tschechien und den USA präsentieren. Naturgemäß kann diese Auswahl die mannigfaltigen Initiativen und Aktivitäten auf dem Feld der Handschriftenforschung nicht vollständig abdecken. Dennoch sucht dieser Band einen umfassenden Überblick über die wichtigsten Ansätze und Strömungen der aktuellen Entwicklung auf den genannten Themenfeldern zu vermitteln.

Die Beiträge sind den traditionellen Arbeitsfeldern der Handschriftenforschung entsprechend zwei großen Bereichen zugeteilt. Für die Kodikologie wird ein Bogen von der Katalogisierung und Erschließung, über die Katalogintegration, die Digitalisierung der Handschriften selbst und die Dokumentation der Wasserzeichen bis hin zur Entwicklung umfassender digitaler Forschungsumgebungen gespannt. Das Feld der Paläographie umfasst die universitäre Lehre, Fragen der Analyse digitaler Abbildungen, Ansätze zu einer Zeichenerkennung handschriftlicher Dokumente, Schriftklassifikation, Schreiberidentifikation und allgemeine Überlegungen zur Weiterentwicklung der Paläographie angesichts der neuen Möglichkeiten.

Das digitale Medium bildet nicht nur den Hintergrund für die hier behandelten Gegenstände. Es bestimmt auch die Produktionsweise eines solchen Sammelbandes. Auf der einen Seite haben die elektronischen Hilfsmittel bei der Verbreitung eines Calls for Papers, bei der Kommunikation zwischen den Autoren und den Herausgebern und bei der kollaborativen Arbeit derselben sehr zur Beschleunigung der Abläufe beigetragen – nur so kann innerhalb von gerade einmal acht Monaten der höchst komplexe Prozess vom Aufruf zur Einreichung bis zur Auslieferung des gedruckten Bandes überhaupt abgewickelt werden. Auf der anderen Seite erhöhen die erweiterten Möglichkeiten und der weitgehende Wegfall verlegerischer Zuarbeit bei Lektorat und Satz den Aufwand enorm: Bis zum fertigen Buch haben auf Seiten des IDE insgesamt zehn Mitglieder und Helfer, auf Seiten der Autoren 38 Forscherinnen und Forscher unzählige Arbeitsstunden investiert; dabei sind ca. 2.000 E-Mails gewechselt, zahlreiche Skype-Chats adhoc angestoßen und ein gutes Dutzend Telefonkonferenzen über den örtlich verstreuten Kreis der Herausgeber und Mitarbeiter abgehalten worden.

Dem internationalen Zuschnitt des Forschungsfeldes entsprechend sind die Beiträge in deutscher, englischer, französischer und italienischer Sprache abgefasst. Allen Beiträgen sind Zusammenfassungen in englisch und deutsch und gegebenenfalls in der Sprache des Aufsatzes vorangestellt. Die bibliographischen Anhänge folgen in ihrer Form weitgehend den Empfehlungen der MLA (Modern Language Association), 6. Auflage. Auf die Angabe des letzten Aufrufs von URLs wurde verzichtet. Hier gilt grundsätzlich: Letzter Besuch 4. Mai 2009.

Die nun vorliegende Anthologie erscheint als zweiter Band in der Reihe der Schriften des Instituts für Dokumentologie und Editorik. Ein Teil der Publikationskosten konnte durch die finanzielle Unterstützung der *Association Paléographique Internationale Culture Écriture Société* (APICES) gedeckt werden. Das IDE dankt dem Moderamen der APICES für die Auszeichnung mit dem erstmals im Angedenken an den Gründungspräsidenten J. M. M. Hermans (1949-2007) verliehenen Preis zur Förderung paläographischer und kodikologischer Forschung. Die Urteilsbegründung lobt neben der thematischen Ausrichtung des Bandes ausdrücklich die innovative Publikationsform, bei der die Veröffentlichung der einzelnen Forschungsbeiträge einem internationalen Symposium zur Diskussion herausragender und zukunftsweisender Ergebnisse vorausgeht. Diese Tagung findet im Juli 2009 an der Ludwig-Maximilian Universität zu München statt. Mit Georg Vogeler (München) konnte ein Mitglied des Moderamen der APICES dafür gewonnen werden, den Band mit einer inhaltlichen Einführung zu versehen.

Allen Beitragenden sei hiermit herzlich für ihre professionelle und zuvorkommende Zusammenarbeit gedankt, die eine schnelle und reibungslose Realisierung dieses Publikationsprojektes ermöglicht hat. Christiane Fritze (Berlin) und Franz Fischer (Dublin) haben die redaktionellen Arbeiten der Herausgeber unterstützt und die Entwicklung des Bandes maßgeblich mitbestimmt. Bernhard Assmann (Köln) hat die Konversion der eingereichten Beiträge nach XeTeX, den Satz und in zahllosen Änderungsumläufen die Einarbeitung von Korrekturen bewältigt. Katharina Weber (Köln) hat in bewährter Manier das Umschlaglayout gestaltet. Gill Bepler (Wolfenbüttel), Doireann Dennehy (Galway) und Katharina Mahler (Köln) danken wir für zusätzliche Korrekturarbeiten.

Vorbemerkung zur elektronischen Fassung

Die elektronische Fassung von »Kodikologie und Paläographie im digitalen Zeitalter« wurde gegenüber der Druckauflage um ein Handschriften-Register erweitert. Außerdem wurden kleinere Korrekturen vorgenommen, und Ligaturen wurden entfernt, um die Volltextsuche zu ermöglichen. Dadurch haben sich an einigen Stellen Verschiebungen in den Seitenumbrüchen ergeben, die sich auf seitengenaue Zitationen auswirken können. Wir haben bei der Erstellung der elektronischen Fassung jedoch sichergestellt, dass das Inhaltsverzeichnis und damit die Start- und Endseiten der Beiträge unverändert geblieben sind. Beiträge können somit mit den gleichen bibliographischen Angaben wie in der Druckfassung referenziert werden.

Würzburg, Köln und Wolfenbüttel, Oktober 2009, die Herausgeber

Preface

We evoke the technological developments which have taken place since the end of the 20th century by speaking of the beginning of a “digital age”. As with other scholarly disciplines, work with manuscripts poses questions about the meaning and importance of this transition to the digital medium and about the influence of the Internet and the omnipresence of computer technology in research. What possibilities are offered by almost limitless storage and computing capacities, advanced scanning technology and image processing, and the Internet as a new central publication platform? What are the chances—and the risks?

Today the products of the “digital age” are already used in many ways. Information resources like manuscript catalogues and watermark databases are digitised and collected in portals: this virtualisation permits new thematic arrangements and the reconstruction of historical collections. The complete digital imaging of manuscripts allows a far simpler access to documents and leads to new modes of using them. With the possibilities of eLearning didactic options for imparting palaeographic reading competence are extended. Finally, quantitative palaeography can be carried out on the basis of a widely available pool of pictures and with the help of computer algorithms. Now letter models and characteristics of writing styles can be exactly determined and stored for reference. This opens up perspectives for computer-aided recognition of scripts and thus new bases for transcriptions and editions which are not cut off from their manuscript sources in publication, but remain closely tied to them.

What is the way forward? How long will it be before we are able to carry out automatic text recognition of medieval manuscripts? Will there be a universal virtual catalogue in the near future which gives access to a completely digitised European or even world-wide manuscript heritage on the Internet? Could collaborative platforms evolve in which historical records are comprehensively described, transcribed and edited? Will research methods be fundamentally changed under these new conditions and lead to new insights about the development of human culture?

Realising old dreams on the one hand and discovering new interrogative horizons on the other: we are in a situation of quick progress and radical change in codicological and palaeographical theory and practise. To document the state of the art and to indicate developing perspectives, the Institute for Documentology and Scholarly Editing (IDE)—a union of young scholars from the area of Humanities Computing—published a Call for Papers for the present anthology at the end of October 2008. The 21 contributions which are published here, presenting work and projects of researchers from Australia, Austria, Czech Republic, France, Germany, Great Britain, Ireland, Italy, and the USA, were selected from a total of 31 submitted proposals. Naturally, this choice cannot completely cover the manifold initiatives and activities in the field of manuscript research. Nevertheless, this volume tries to provide an overview of the most important developments and current projects.

The articles are assigned according to the traditional fields of work in manuscript research. From the area of codicology, topics range from cataloguing and documentation, integrative catalogues, the digitisation of manuscripts, and the documentation of watermarks to the creation of all-embracing digital research environments. The field of palaeography encompasses university teaching, issues of digital image analysis, approaches to character recognition of hand-written documents, classification of script, identification of individual hands and general considerations of the enhancement of palaeography.

The digital medium does not just form the background for the objects treated here. It also determines the process of producing such a volume. On one hand, electronic media were used to spread the Call for Papers and they provided the basis of communication between authors and editors and accelerated the workflow between them. This was the only way in which the extremely complicated process from the call for papers, to the submission, up to the production of the printed volume could have been handled within just eight months. On the other hand, the possibilities which electronic media offer bring enormous savings in time and costs for proof-readers and typesetting. A team of ten members and assistants from the IDE and 38 researchers invested countless working hours; approximately 2,000 emails were exchanged and numerous ad-hoc Skype chats and a dozen phone conferences with the widely scattered circle of the editors and colleagues were held.

In keeping with the international orientation of the research field, the articles are written in English, French, German, and Italian. Summaries in English and German and, if necessary, in the language of the article, precede all contributions. The bibliographical appendices broadly follow the recommendations of the MLA (Modern Language Association), 6th edition. We omitted the quotation of the last check on URLs. In all cases it can be assumed that the last visit of each site was on 4th of May, 2009.

The present volume is published as the second in a series of publications of the Institute for Documentology and Scholarly Editing (Schriften des Instituts für Dokumentologie und Editorik). Part of the publication costs was covered by the financial support of the *Association Paléographique International Culture Écriture Société* (APICES). The IDE is grateful for receiving the first award for the promotion of palaeographical and codicological research from the Moderamen of the APICES which has been established in honour of the founding president J. M. M. Hermans (1949–2007). Besides praising the content of the volume and its thematic approach, the award statement explicitly mentions the innovative form of the publication which precedes an international symposium to discuss the most outstanding and visionary results. This conference takes place in July 2009 at the Ludwig-Maximilian University in Munich. Georg Vogeler

(Munich), a member of the Moderamen of the APICES, was kind enough to provide the volume with a scholarly introduction.

We want to thank all contributors for their professional co-operation which made the quick and smooth realisation of this project possible. Christiane Fritze (Berlin) and Franz Fischer (Dublin) have supported the editorial work of the editors and had a significant influence on the creation of the volume. Bernhard Assmann (Cologne) mastered the conversion of the submitted contributions to XƎTEX, the typesetting and the incorporation of corrections in countless proof circulations. Katharina Weber (Cologne) designed the cover. We thank Gill Bepler (Wolfenbüttel), Doireann Dennehy (Galway), and Katharina Mahler (Cologne) for additional proof-reading.

Notes on the Electronic Version

The electronic version of “Codicology and Palaeography in the Digital Age“ contains an additional index of manuscripts. Furthermore, it has been slightly edited in comparison to the print, some corrections have been made and ligatures eliminated to allow fulltext search. This has caused minor changes in page breaking which might have some impact on citations. We have, however, assured that the table of contents, i.e. start and end pages of all articles, has remained unchanged. Hence, the same bibliographic details for referencing articles as for the printed version can be used.

Würzburg, Cologne and Wolfenbüttel, October 2009, the editors

Einleitung

Der Computer und die Handschriften

Zwischen digitaler Reproduktion und maschinengestützter Forschung

Georg Vogeler

Zusammenfassung

Die Beiträge dieses Bandes zeigen, dass sich die Diskussion über den Computer als Medium für die Verbreitung von Grundlagen und Ergebnissen der Forschung nicht losgelöst von einer Diskussion über den Computer als Forschungsinstrument führen lässt. Die digitale Reproduktion von Handschriften determiniert die Art und Weise ihrer Erforschung. Einerseits verändern sich Verarbeitung und Beschreibung eines handschriftlichen Zeugnisses; andererseits schließt sich an seine digitale Repräsentation die Forschungsdiskussion im Medium des Internets unmittelbar an. Die Einleitung versucht aus diesem Sachverhalt Schlussfolgerungen auf mögliche Zukunftsszenarien einer »digitalen Kodikologie« und einer »digitalen Paläographie« abzuleiten.

Abstract

The papers in this anthology show that the distinction of computer technology between a medium for distribution of materials for teaching and research on the one hand and an instrument for research on the other is common but not sensible. The digital representation of manuscripts determines scholarly work: the way manuscripts are being described is changing, while digital representation acts as a starting point for an academic discussion within the digital medium itself. This introduction attempts to map out possible future horizons for "digital palaeography" and "digital codicology".

Die Association Paléographique Internationale – Culture, Écriture, Société (APICES) hat sich neben der Comission Internationale de Paléographie Latine (CIPL) seit 1995 als zweite wichtige internationale Vereinigung von Wissenschaftlerinnen und Wissenschaftlern etabliert, deren Forschungsinteresse auf Handschriften ausgerichtet ist. Das Moderamen der APICES unterstützt aktiv moderne paläographische und kodikologische Forschungen und hat sich deshalb an diesem Band beteiligt, nicht nur finanziell, sondern auch bei der Kür von vier Beiträgen, die als besonders instruktiv und richtungsweisend ausgezeichnet wurden. Diese Wahl fiel sehr schwer. Dennoch konnte sich das Moderamen mit den Herausgebern des Bandes darauf einigen, dass die Beiträge von

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. XV–XXIV.

Wernfried Hofmeister, Andrea Hofmeister-Winter und Georg Thallinger, von Timothy Stinson, von Peter Stokes sowie von Roland und Gilbert Tomasi eine Hervorhebung in dieser Hinsicht verdienen. Dass sich dem Paläographen und Kodikologen die anderen Beiträge ebenso gut in das Gesamtbild einer angeregten und Perspektiven öffnenden Forschungsentwicklung fügen, soll diese Einleitung zeigen.

Wie könnte dieses Gesamtbild aussehen? Der Einsatz des Computers für geisteswissenschaftliche Forschung wird unter zweierlei Aspekten diskutiert, und die Teilnehmer an den beiden Diskussionen kommunizieren erstaunlich wenig miteinander. Die eine Diskussion geht von den Möglichkeiten aus, die Computernetzwerke als Medium, d.h. als Publikations- und Informationsraum der Geisteswissenschaften, eröffnen. Für Paläographie und Kodikologie geht es in diesem Bereich um die Digitalisierung von Handschriften und ihrer Beschreibungen, um Online-Kataloge und die richtigen Parameter, mit denen digitale Bilder von den Handschriften erstellt werden sollen, um das Internet als Publikationsort von Forschungsergebnissen oder als paläographischen Lernort und schließlich auch um das Internet als Kommunikationsraum des wissenschaftlichen Diskurses.

Auf dem anderen Diskussionsfeld geht es um die Möglichkeiten und Probleme, mit Hilfe des Computers alte Forschungsfragen endlich auf eine befriedigende Weise zu beantworten oder aber neue Fragen aufzuwerfen, deren Beantwortung mit einer herkömmlichen Methodik nicht nur unmöglich gewesen, sondern die zu stellen uns gar nicht in den Sinn gekommen wäre. Hier geht es um die Aussagekraft von Maßen und quantitativen Daten für die Analyse und Beschreibung von mehr oder weniger kalligraphischen Produkten mittelalterlicher Schreiber, um kodikologische Statistik, um die Erkennung von Regelmäßigkeiten in Zeichenrepositorien ausgewählter Handschriftengruppen oder um die Transkription von Handschriften als Vorarbeiten umfassender, kritischer Editionen, um nur einige der denkbaren Anwendungsszenarien zu benennen.

Die im vorliegenden Band zusammengeführten Projektberichte und Forschungsergebnisse lassen sich denn diesen beiden Diskussionen zuordnen: Der Beitrag von *Bernard Muir* zum »Ductus«-Projekt beschreibt die Geschichte eines computergestützen Lernhilfsmittels, dessen Besonderheit darin liegt, die aktuellen multimedialen, technischen Möglichkeiten voll auszuschöpfen. *Marco Palma* und *Antonio Cartelli* diskutieren, wie sich das Lernverhalten von Studenten mit zunehmender Fülle an digitalem Unterrichtsmaterial ändert resp. verschlechtert und wie das Medium »Internet« als Kommunikationsplattform dienen kann, in der die Studierenden mit einem von den Autoren entwickelten Online-Informationssystem für die paläographische Lehre kollaborativ paläographisches Wissen erwerben, indem sie aus einer Beispielsammlung eigenständig Kurse zusammenstellen. Für *Silke Kamp* ist der Medienwandel vom gedruckten Buch zur interaktiven Webseite der Schlüssel für eine neue paläographische Fachdidaktik.

Die Möglichkeiten, Informationen über Handschriften schneller und leichter zu transportieren, werden auch in vielen kodikologischen Beiträgen dieses Bandes abgehandelt: *Christina Wolf* spricht z.B. über das Internet als Medium für eine Datensammlung von Wasserzeichen. Das Netz ermöglicht es, Kataloge oder Bilder von Handschriften einer Region an einem virtuellen Ort zusammenzufügen und zugänglich zu machen, und eröffnet damit neue Wege, das in den Bibliotheken bereits vorhandene oder noch zu erarbeitende Wissen über Handschriften zu kommunizieren. Um dasselbe Themenfeld kreisen die Beiträge zur Online-Präsentation der Handschriften der Universitätsbibliothek Heidelberg von *Pamela Kalning* und *Karin Zimmermann*, zum regionalen Handschriftenportal des Veneto von *Francesco Bernardi*, *Paolo Eleuteri* und *Barbara Vanin* und zum »Offenen Katalog« der Biblioteca Malatestiana in Cesena von *Paola Errani*, *Antonio Cartelli*, *Andrea Daltri*, *Marco Palma* und *Paolo Zanfini*. Hinzu kommen die Beiträge zum Aufbau eines europäischen Handschriftenkatalogs von *Zdeněk Uhlíř* und *Adolf Knoll* sowie zur virtuellen Rekonstruktion des handschriftlichen Nachlasses des Luthermitarbeiters Georg Rörers von *Christian Speer*. Die Rückwirkungen des digitalen Mediums auf die sich wandelnde Praxis der Handschriftenbeschreibung diskutiert *Timothy Stinson*.

Einige der Beiträger sehen im Computer und seiner Vernetzung mehr als nur eine verbesserte Präsentationsform der Handschriften und ihrer Erschließungshilfsmittel. Inspiriert von der Diskussion über das »Web 2.0« (»Social Web«) erscheint das Internet nicht nur ein Medium zur einseitigen Verbreitung von Informationen, sondern auch und in zunehmendem Maße als ein Ort des Austausches und der gemeinsamen Arbeit. Webseiten mit Handschriftenkatalogen können dann Orte sein, an denen sich Wissenschaftler über die Handschriften einer Bibliothek austauschen und die Kataloge fortlaufend durch eigene Forschungsergebnisse bereichern. *Peter Stokes* spitzt diese Überlegung so weit zu, dass man den Eindruck bekommt, nachprüfbare paläographische Forschung sei nur über die Veröffentlichung ihrer Methoden in dafür eingerichteten Angeboten des »Social Web« möglich, in denen die Paläographen nicht nur ihre schwer bis gar nicht nachvollziehbare Expertise publizieren, sondern ebenso die Messdaten und Berechnungsmethoden, auf denen sie beruhen. Das Internet als Medium der paläographischen und kodikologischen Forschung ist in diesem Diskussionsbereich zum Instrument geworden, das Argumente über die handschriftlichen Produkte des Mittelalters und der Frühen Neuzeit nachvollziehbar macht und damit wissenschaftliche Diskussion überhaupt erst ermöglicht.

Der zweite Diskussionsstrang, um den sich die Beiträge dieses Bandes gruppieren lassen, konzentriert sich auf die wissenschaftliche Arbeit, die geleistet wird, nachdem die Forscher die Objekte für ihre Arbeit ausgewählt haben und bevor sie ihre Forschungsergebnisse publizieren. Die Diskussion um den Einsatz des Computers bei der Erforschung der handschriftlichen Zeugnisse selbst zielt derzeit vor allem auf das »Paläographische Messen«. Der Beitrag von *Maria Gurrado* stellt eine Erweiterung für ein

Open-Source-System vor, mit dessen Hilfe die wichtigsten Maße eines Schriftbeispiels ermittelt werden können. *Mark Aussems* und *Axel Brink* suchen ebenso nach aussagekräftigen Maßzahlen zur Schreiberidentifikation wie *Wernfried Hofmeister*, *Andrea Hofmeister-Winter* und *Georg Thallinger*, wobei die einen die Buchstaben selber ausmessen, während die anderen eine Statistik von händisch ermittelten Befunden mit Mustererkennung kombinieren.

Die computergestützte Ermittlung von schreiberübergreifenden Merkmalen von Schrift ist das Thema von Arianna Ciula wie auch von Mark Stansbury. *Arianna Ciula* sucht nach den Gemeinsamkeiten von automatisch ermittelten Buchstabenmodellen. *Mark Stansbury* zeigt auf, dass der Computer mit seinen Möglichkeiten zur Verarbeitung großer Mengen von Handschriften nicht ausschließlich einer rein systematisierenden Herangehensweise an die Paläographie Vorschub leisten wird, sondern auch dem nach der Evolution der Schriftarten fragenden Forschungsansatz aufschlussreiche Ergebnisse liefern kann.

Der Beitrag von *Gilbert* und *Roland Tomasi* verknüpft die Suche nach Maßzahlen zur Schreiberidentifikation mit den geometrischen Informationen, die bei der automatischen Schrifterkennung ermittelt werden. Er verbindet damit das Vermessen der Schrift mit einem dritten Bereich, in dem der Computer auf handschriftliche Zeugnisse angewendet wird: die computergestützte Transkription. *Daniele Fusi* diskutiert das Konzept lernender neuronaler Netzwerke als einer möglichen Grundlage für die automatische Erkennung von handschriftlichen Texten. *Hugh Cayless* beschreibt Möglichkeiten der Verknüpfung des Textes mit dem Bild der Handschrift. Digitale Kodikologie und digitale Paläographie zielen also auch auf die Transkription der Texte und auf die dauerhafte Verknüpfung der destillierten linguistischen Codes mit der Visualität der ihnen zugrunde liegenden realen Dokumente.

Paläographische und kodikologische Forschung mit dem Computer ist jedoch primär noch eine messende Wissenschaft. Die an den Formen gewonnenen Messdaten aber bergen so umfangreiche Deutungsmöglichkeiten, dass deren Gewichtung immer auch die Expertise des Handschriftenforschers erfordert. Dessen mehr oder weniger intuitive Urteile stehen in produktiver Konkurrenz zu den computergestützten Methoden und werden in den hier vorgestellten Forschungsansätzen immer als letztlich entscheidendes Korrektiv benötigt.

Mit einer Einordnung in die Diskussionsstränge der messenden Forschung an den Schriftzeugnissen einerseits sowie der medialen Vermittlung von Handschriften und dem Wissen darüber andererseits ist aber noch längst nicht alles über die hier vorgelegten Beiträge gesagt. Schon der Beitrag von Peter Stokes zeigt, dass eine neue mediale Umgebung die messende paläographische Forschung befördert, indem sie eine Plattform bietet, um sich über die den Messergebnissen zu Grunde liegenden Daten und Methoden auszutauschen. Die Zusammenschau der Beiträge zeigt, dass eine Trennung der Diskussionen keinen Sinn macht und eröffnet dagegen den Blick auf neue Forschungsperspektiven: Die Handschriften, die Arianna Ciula, Wernfried Hofmeister und seine Mitarbeiter, Mark Aussems und Axel Brink ausgemessen haben, sind zunächst einmal digitalisiert worden, und zwar nur für den jeweils spezifischen Forschungszweck. Die Arbeitsbedingungen für solche Forschungen ändern sich aber fundamental mit der anwachsenden Verbreitung von online vermittelten Bildern von Handschriften, in den Webseiten der Malatestiana, der Universitätsbibliothek Heidelberg, der Kölner Diözesan- und Dombibliothek (CEEC), der Sankt Galler Handschriften (CESG), der Herzog August Bibliothek (HAB) oder der Bayerischen Staatsbibliothek (BSB): Als erstes lösen sich die Handschriften von ihrem physikalischen Aufbewahrungsort. So können handschriftliche Zeugnisse Georg Rörers gemeinsam auf einer Oberfläche untersucht werden, auch wenn sie in verschiedenen Bibliotheken liegen. Darüber hinaus sind die elektronischen Bilder der Handschriften auch Objekt von messenden Methoden: Die Techniken der Mustererkennung, die bei den auf der Vorarbeit von Lambert Schomaker beruhenden Versuchen von Mark Aussems und Axel Brink angewendet werden, ebenso wie die Strukturerkennungen, auf denen die Text-Bild-Verknüpfungen von Hugh Cayless oder Gilbert und Roland Tomasi beruhen, können auf beliebige Handschriftenbilder angewendet werden. Aus der Präsentation des Handschriftenbestandes einer Bibliothek wird also ein Baustein für eine weit ausgreifende computergestützte Forschung.

Für die Kodikologie sind Kataloge, die den Existenznachweis von Handschriften führen, grundlegend. Die Vorstellung, das Medium Computer als Findmittel zu benutzen, das die Existenz von Handschriften über die Grenze einer Bibliothek hinaus, ja sogar unter bewusster Verwischung von modernen Bibliotheksgrenzen, als Rekonstruktion verlorener Bibliotheken nachweisen kann, hat Ezio Ornato zu Visionen angeregt, die angesichts der in diesem Band vorgestellten Forschungsprojekte ein Stück weit realisierbarer erscheinen. Die Rekonstruktion der berühmten Bibliothek des Humanistenkönigs Mathias Corvinus von Ungarn in der Bibliotheca Corviniana Digitalis mag als Beispiel dienen, wie der Nachweis der Handschriften einer solchen virtuellen Bibliothek aussehen kann. Der vorliegende Band bietet mit den Beiträgen von *Paola Errani*, *Antonio Cartelli*, *Andrea Daltri*, *Marco Palma* und *Paolo Zanfini* zum »Offenen Katalog« der Handschriften der Biblioteca Malatestiana in Cesena (Emilia-Romana) und von *Franceso Bernardi*, *Paolo Eleuteri* und *Barbara Vanin* zur »Nuova Biblioteca Manoscritta« zwei weitere Beispiele dafür, welche Aufgaben zu bewältigen sind und welches Potential in einem weiter ausgreifenden elektronischen Bestandsnachweis liegen kann. Die Projektbeschreibung zur »Nuova Biblioteca Manoscritta« deutet an, mit welchen Detailproblemen die Arbeit in verschiedenen Bibliotheken in einem Katalog zusammengeführt werden kann. Demgegenüber erscheint die Perspektive eines auf reinem »Harvesting« beruhenden gesamteuropäischen Handschriftenkatalogs, wie sie *Zdeněk Uhlíř* und *Adolf Knoll* als ein Ziel des ENRICH-Projektes aufmachen, bei aller Attraktivität derzeit noch riskant. Die Lösungen, die der Beitrag für die mehrsprachige und auf Inhalten beruhende Suche andeutet, sind noch nicht realisiert.

Timothy Stinson denkt diese Entwicklung bis zu dem Punkt weiter, an dem die etablierten Kategorien der Handschriftenbeschreibung durch ihre digitale Abbildung und ihre Verflechtung mit digitalen Bildern durch neuen Kategorien zu ersetzen sind. Damit eröffnen sich Perspektiven, die auch auf einen etablierten Bereich der Computernutzung in der Kodikologie zurückwirken können: Seit den 1980er Jahren werden die Handschriftenkataloge als Teil einer Archäologie der Handschrift mit interessanten Ergebnissen statistisch ausgewertet (Ornato 1997, Maniaci). Diese zählen die Kategorien aus, die in den Handschriftenkatalogen verwendet werden. Es mutet wahrscheinlich an, dass sich aus einer Digitalisierung der Kataloge – und das meint hier nicht nur die digitale Reproduktion einer Druckseite, sondern auch die Kodierung ihrer Inhalte – neue Fragestellung für diese Forschungsrichtung ergeben können.

Dass das Katalogisat einer Handschrift in vielen Einzelfacetten Forschungspotential bietet, zeigt der Beitrag von *Christina Wolf*. Die von ihr vorgestellten Überlegungen zum Aufbau einer Wasserzeichendatenbank in Fortführung des Projektes Bernstein ist zunächst ein Nachweisinstrument. Sie ist aber ebenso ein Hilfsmittel zur Forschung, in dem kollaborativ Daten zusammengetragen werden. Allerdings scheint dieses System mit seinem geschlossenen Kreis an Zuträgern sogar noch hinter den Möglichkeiten einer offeneren Zusammenarbeit zurückzubleiben, wie sie der »Offene Katalog« der Biblioteca Malatestiana als Möglichkeit kollaborativer Forschung auf einem eigens eingerichteten Forum und sog. *cantieri* (»Baustellen «) bereits mit einigem Erfolg erprobt.

Nicht alle Facetten der Forschungen an Handschriften haben einen Niederschlag in diesem Sammelband finden können. So fehlen Beiträge zur Kunstgeschichte der Handschrift oder zu Notenhandschriften. Hier wären unter anderem die Versuche von Manuscripta Mediaevalia oder der UB Heidelberg zu reflektieren, mit Hilfe von IconClass eine mehrsprachige Suche nach Bildinhalten und Schlagwörtern zu ermöglichen und eine Art Ontologie der bildlichen Darstellungen in die Online-Präsentation von Handschriften zu integrieren. Das Verhältnis zwischen der wachsenden Zahl an digital reproduzierten Handschriften und dem gleichzeitigen Wachsen computergestützter Forschung an diesen Bildern lässt die Vermutung zu, dass der Mangel an Beiträgen zu computergestützten kunsthistorischen Forschungen an mittelalterlichen Handschriften dem Umstand geschuldet ist, dass das entsprechende Bildmaterial erst in geringer Menge digital verfügbar ist.

Die im Band vorgestellten Beiträge aus dem Bereich der Kodikologie beschränken sich natürlich nicht auf IT-Systeme zum Nachweis von Handschriften oder einzelnen Facetten einer Handschrift. Die im Vergleich zu den gedruckten Faksimiles unvergleichlich geringen Kosten der Reproduktion von digitalen Bildern haben einige Bibliotheken motiviert, Bilddigitalisate ganzer Handschriften im Netz zugänglich zu machen, ja das ENRICH-Projekt versteht sich selbst als Vorstufe zur Handschriftenabteilung der virtuellen europäischen Bibliothek »Europeana«. Sie werden aber auch zu virtuellen Treffpunkten, wenn sie eine Plattform für die Kommunikation der Benutzer anbieten. So kann sich das Projekt der Digitalisierung des handschriftlichen Materials zum Reformator Georg Rörer als Knotenpunkt vielfältiger Forschungsaktivitäten zur Reformationsgeschichte verstehen. Von mannigfachen Forschungsinteressen angetrieben ist deshalb auch TEUCHOS, das von *Daniel Deckers*, *Lutz Koch* und *Cristina Vertan* in diesem Band vorgestellt wird und sowohl kodikologische als auch philologisch-editorische Informationen über Handschriften integriert.

Die Beiträge zur Kodikologie zeigen damit einen Weg auf, den die neuen Informationstechnologien für die Arbeit mit Handschriften ermöglichen: den Weg vom traditionellen bibliotheksbezogenen Nachweis zu integrierten Systemen, in denen sich die Handschriften in ihrer Vielfalt abbilden, als Textträger ebenso wie als physikalische Gegenstände, als Spuren eines Schreibers und Sammlers ebenso wie als gemeineuropäisches Handschriftenerbe, über das sich eine in der ganzen Welt verstreute Gruppe von Wissenschaftlerinnen und Wissenschaftlern im Internet in zwar virtueller aber großer Nähe zu ihrem Objekt austauscht. Dabei werden, wie *Timothy Stinson* betont, die Kategorien der Handschriftenbeschreibung mit neuen, erweiterten Inhalten gefüllt und mit neuen Funktionen aufgeladen – wo nicht durch die Bedingungen digitaler Repräsentation gar völlig neue Beschreibungselemente entstehen.

Auch das Verhältnis von moderner Typographie und Paläographie ist durch den Computer verändert worden. Wie Marc Smith in seinem jüngsten Beitrag (2008) in der GLM beobachtet hat, ermöglicht der Computer die Übernahme von historischen Schriftentwürfen in moderne Druckverfahren, was den Entwurf von Typen nach historischen Vorbildern erleichtert. Die Medieval Unicode Font Initiative (MUFI) versucht umgekehrt, paläographische Phänomene in den Unicode-Zeichensatz zu integrieren und den Entwurf von Typen zu unterstützen, die diese Zeichen darstellen.

In der sich dagegen auf den Umgang mit historischen Schriften konzentrierenden Paläographie hat der Computer als Medium insbesondere im Bereich des Wissenserwerbs zu neuen Konzepten angeregt: Die Beobachtung, dass photographische Reproduktionen leicht, rund um die Uhr und von überall auf der Welt einsehbar werden, hat verschiedene Paläographen daran denken lassen, das traditionelle Tafelwerk als Unterrichtsmaterial im Netz abzubilden. *Bernard Muir* berichtet anschaulich, wie die sich ausweitenden multimedialen Möglichkeiten zu Produkten führen, die sich von einer photographischen Reproduktion der Handschrift, begleitet von Transkription und Erläuterung deutlich entfernen. *Silke Kamp* sowie *Marco Palma* und *Antonio Cartelli* ziehen aus ihren Erfahrungen mit eLearning Schlüsse für didaktische Überlegungen: die eine beschreibt Methoden der Präsenzlehre, die sich in eLearning-Umgebungen noch nicht wiederfinden, aber wiederfinden könnten; die anderen betonen die Möglichkeiten, die eLearning für kollaboratives Lernen bieten.

Ein wichtiges Ziel des paläographischen eLearning ist die Fähigkeit, Texte zu entziffern. Es leuchtet deshalb ein, dass sich ein Forschungsstrang des Einsatzes der IT in der Paläographie mit der Frage beschäftigt, wie aus Bildern Texte werden. Der Vorschlag von *Hugh Cayless* nutzt im Anschluss daran das XML-basierte Vektorgraphikformat SVG, um das Bild gewissermaßen als Text zu beschreiben und so direkt mit der Transkription zu verschmelzen. *Patrick Shiel*, *Malte Rehbein* und *John Keating* nutzen hyperspektrale Scan-Verfahren bei der Bilderstellung, um durch Algorithmen verborgenen Text sichtbar zu machen und Text zu segmentieren und für die Transkription vorzubereiten.

Gilbert und *Roland Tomasi* weisen darauf hin, dass die bei einer automatischen Texterkennung angewendeten Verfahren auch Daten zur Identifikation von Schreiberhänden liefern. Diese Methoden sind der zweite Bereich, in dem computerbasierte Methoden in der Paläographie eingesetzt werden. Während Strukturelemente der Schrift mit der Software von Tomasi automatisch erkannt werden, hat *Maria Gurrado* ältere Ansätze wie die von Patrick Sahle aufgegriffen, ein Hilfsmittel zu entwickeln, das den Paläographen dabei unterstützt, Merkmale wie Schriftwinkel oder Proportionen auszumessen. Die für systematische Rückschlüsse geeigneten Maßzahlen sind es, die auch *Mark Aussems* und *Axel Brink* beschäftigen – und sie stellen den traditionellen Maßzahlen, die eine Automatisierung des geschulten Paläographenblicks sein wollen, mit der durchschnittlichen Strichbreite abstrakte Maßzahlen zur Seite, die sich bei einem nur auf das Visuelle gestützten Urteil nicht berücksichtigen ließen. *Wernfried Hofmeister*, *Andrea Hofmeister-Winter* und *Georg Thallinger* bauen eine Datenbank auf, um mit einer Kombination von graphetischer Statistik, Mustererkennung und kodikologischen Befunden Schreiber erkennen zu können. Die semiautomatische Klassifizierung von Schrift und Schriftzeichen ist das Ziel, das Daniele Fusi, Mark Stansbury und Arianna Ciula bewogen hat, die Nutzbarkeit des Computers für paläographische Analysen zu erproben.

Die Beiträge können zeigen, dass das Messen auch in der Paläographie eine sinnvolle und nützliche Methode ist. *Arianna Ciula* versucht die paläographischen Forschungsmethoden in einem halbautomatischen System abzubilden, in dem graphische Modelle der Buchstaben gebildet und geordnet werden. *Peter Stokes* ergänzt das hier vorgestellte Methodenrepertoire durch die für den forensischen Schriftvergleich genutzten Verfahren, die einer juristischen Prüfung standhalten müssen, und stellt die Kriterien der Nachprüfbarkeit der neuen Methoden denen traditioneller Methoden gegenüber, die sich vor allem auf Autoritäten stützten. Der Computer als Medium kommt genau hier zum Tragen, indem er die Nachprüfbarkeit der neuen Forschungsmethoden mit Hilfe eines Repositoriums paläographischer Daten und einer zu ihrer Analyse verwendeten Software gewährleisten kann. Stokes zeigt damit, dass die Integration des Computers als Forschungsinstrument in seiner Funktion als Medium zu einer produktiven Verwendung von Informationstechnologie führen kann, die sich auf eine kritische Auseinandersetzung mit neuen Methoden stützt und so das Urteilsvermögen des geschulten Auges an die Evidenzen paläographischer Messwerte zurückbinden lässt.

Wenn man so den Stand der Dinge zusammenfasst, dann ergeben sich in der Tat neue Perspektiven für eine Forschung im Bereich der Kodikologie und der Paläographie im digitalen Zeitalter: Die Kodikologie muss zunächst die reiche, aber nicht normierte Beschreibung der Handschriftenkataloge in computergestützt auswertbare Konzepte zusammenfassen, die im Sinne von Stinson die Beschreibung und die Repräsentation der Handschrift im Bild soweit integrieren, dass Fragen über das reine Auszählen einzelner Beschreibungsparameter hinaus möglich werden. Dazu bieten der von Denis Muzerelle initiierte »Vocabulaire Codicologique« ebenso Ansätze wie übergreifende, integrierende Kataloge oder die von Hugh Cayless für die Transkription angerissenen Überlegungen, Text und Bild miteinander zu verschmelzen und als ein Objekt digital zu repräsentieren. Dabei wird Mustererkennung eine wichtige Rolle spielen, wobei noch nicht klar ist, wie die entsprechenden Muster zu errechnen sind, denn die Aussagekraft der messbaren Eigenschaften einer Schrift ist bei weitem noch nicht ausreichend erforscht. Während Melissa Terras 2006 erfolgreich die paläographische Arbeit bei der Entzifferung der »Tabulae Vindolanda« mit einem Computersystem unterstützt hat, das auf einer detaillierten Analyse des vom Paläographen angewendeten Leseprozesses beruht, zeigen die Überlegungen von Daniele Fusi ebenso wie die Ergebnisse von Mark Aussems und Axel Brink, dass man auch für den Paläographen zunächst fremde Methoden ausprobieren darf und testen muss, um ihre Aussagekraft mit den traditionellen Methoden zu vergleichen. Eine »digitale Paläographie« und eine »digitale Kodikologie« als Selbstverständnis einer Gruppe von Handschriftenforschern kann dazu den sozialen Hintergrund bilden, der einen kreativen und produktiven Austausch ermöglicht, sei es in gedruckten Publikationen, im »Social Web« oder in der persönlichen Kommunikation auf Tagungen und in den Handschriftenbibliotheken.

Bibliographie*

APICES: *Association Paléographique Internationale – Culture, Écriture, Société.* <http://www.palaeographie.org/apices/>

Bibliotheca Corviniana Digitalis. Virtual reconstruction of King Matthias' Library. <http://www.corvina.oszk.hu>

BSB: *Bayerische Staatsbibliothek. Münchner Digitalisierungszentrum.* <http://www.digitale-sammlungen.de>

CEEC: *Codices Electronici Ecclesiastici Coloniensis.* <http://www.ceec.uni-koeln.de>

CESG: *Codices Electronici Sangallenses.* <http://www.cesg.unifr.ch>

* Die im Text zitierten Beiträge dieses Bandes sind nicht aufgenommen.

CIPL: *Comission Internationale de Paléographie Latine.* <http://www.palaeographia.org/cipl>

Europeana. < http://www.europeana.eu/portal>

HAB: *Herzog August Bibliothek Wolfenbüttel. Digitalisierte Handschriften, Sondersammlungen.* <http://www.hab.de/bibliothek/wdb/mssdigital.htm>

IconClass. <http://www.iconclass.nl>

IconClass Browser. <http://www.iconclass.nl/libertas/ic?style=index.xsl>

Maniaci, Marilena. *Archeologia del manoscritto. Metodi, problemi, bibliografia recente.* Con contributi di Carlo Federici e di Ezio Ornato. I libri di Viella, 34. Rom: Viella, 2002.

Manuscripta Mediaevalia. <http://www.manuscripta-mediaevalia.de>

MUFI: *Medieval Unicode Font Initiative.* <http://www.mufi.info>

Muzerelle, Denis, ed. *Vocabulaire Codicologique, Répertoire méthodique des termes français relatifs aux manuscrits avec leurs équivalents en anglais, italien, espagnol.* Version 1.1 2002-2003 <http://vocabulaire.irht.cnrs.fr/vocab.htm>

Ornato, Ezio. »L'historie du livre et les méthodes quantitatives: bilan de vingt ans de recherches.« *La face cachée du livre mèdiéval. L'histoire du livre vue par Ezio Ornato, ses amis et ses collègues.* I libri di Viella, 10. Viella: Roma 1997: 607-679.

Ornato, Ezio. »La codicologie quantitative, outil privilégié de l'histoire du livre médiéval.« *La face cachée du livre mèdiéval. L'histoire du livre vue par Ezio Ornato, ses amis et ses collègues.* I libri di Viella, 10. Viella: Roma 1997 375-472.

Ornato, Ezio. »Bibliotheca manuscripta universalis. Digitalizzazione e catalografia: un viaggio nel regno dell'utopia?« *Gazette du livre médiéval* 25 (2006). 1-13. <http://www.palaeographia.org/glm/glm.htm?art=utopia>

Sahle, Patrick: *Werkzeug zur paläographischen Dokumentation von Handschriften.* <http://www.ceec.uni-koeln.de/projekte/CEEC/tools/paleography/paleography.htm>

Smith, Marc H. »Du manuscrit à la typographie numérique: présent et avenir des écritures anciennes.« *Gazelle du livre Médiéval* 52-53 (2008): 51-78.

Terras, Melissa. *Image to Interpretation. Intelligent Systems to Aid Historians in the Reading of the Vindolanda Texts.* Oxford Studies in Ancient Documents. Oxford: Oxford University Press, 2006.

Kodikologie: Vom Katalog zur Forschungsumgebung

Codicology: From Catalogue to Virtual Research Environment

La catalogazione in rete dei manoscritti delle biblioteche venete: *Nuova Biblioteca Manoscritta*[*]

Francesco Bernardi, Paolo Eleuteri, Barbara Vanin

Riassunto

Nuova Biblioteca Manoscritta (NBM) è il catalogo in linea dei manoscritti conservati nelle biblioteche del Veneto – stimati in ca. 90.000, non tenendo conto dei carteggi – senza limitazioni cronologiche o di contenuto. Questo patrimonio è fino ad oggi accessibile in maniera incompleta e insufficiente mediante cataloghi a stampa parziali, spesso per di più poco rispondenti alle esigenze scientifiche moderne. Il progetto, finanziato dalla Regione del Veneto, è iniziato nel 2003 e vi partecipano attualmente 38 biblioteche. Il lavoro di catalogazione, che privilegia in generale una descrizione di tipo sommario, si svolge via Internet attraverso la catalogazione partecipata di più biblioteche, che lavorano sulla stessa banca dati. I catalogatori condividono in rete le liste di autorità dei nomi, dei titoli, degli argomenti, delle antiche segnature, della tipologia del testo e del genere letterario, della bibliografia; si ha così il vantaggio di accedere a informazioni già strutturate e di poter aggiornare continuamente le notizie, nello spirito proprio di un catalogo aperto. Tutta la gestione di NBM si svolge attraverso Internet, dalla catalogazione sino alla revisione delle schede e alla pubblicazione finale, secondo diversi profili che corrispondono alle differenti funzioni nell'ambito del progetto. Un coordinamento scientifico provvede al controllo e alla revisione di ogni scheda descrittiva, all'assegnazione delle chiavi di accesso all'area di catalogazione, alla gestione dei contenuti del sito. Per garantire la maggiore uniformità possibile nelle descrizioni sono state elaborate delle linee guida per la catalogazione. In NBM è possibile allegare immagini relative ad ogni parte della scheda di descrizione, ma anche importare materiale digitalizzato integralmente, consentendone una consultazione pagina per pagina. L'interrogazione della banca dati di NBM è possibile attraverso l'OPAC presente sul sito e mediante il protocollo Z39.50. Fino ad oggi i manoscritti catalogati, pubblicati e consultabili sono più di 19.000.

Zusammenfassung

Die *Nuova Biblioteca Manoscritta* (NBM) ist ein OPAC der Handschriften der Bibliotheken des Veneto. Er beschreibt ca. 90.000 Objekte ohne zeitliche und inhaltliche Be-

[*] Le versioni precedenti del progetto sono state già presentate altrove: cfr. Eleuteri e Vanin 2005, Vanin e Eleuteri 2006, Eleuteri e Vanin 2007, Eleuteri 2007, Vanin 2008.

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 3–11.

schränkung. Gedruckte Kataloge können dieses Kulturerbe nur unvollständig abbilden und erfüllen darüber hinaus auch nicht die Anforderungen moderner Forschung. Das hier vorgestellte Projekt, finanziert von der Region Veneto, will diese Nachteile überwinden. Das Projekt hat 2003 begonnen und enthält inzwischen Material aus 38 Bibliotheken. NBM ist Internet-basiert. Sein Kern ist eine zentrale Datenbank, in die teilnehmende Bibliotheken ihre Daten einspeisen. Die Handschriftenbearbeiter nutzen zentrale Thesauri für Namen, Titel, Schlagwörter, alte Signaturen, bibliographische Angaben, Textarten und Genres. Sie haben Zugriff auf die schon eingespeisten Daten, die sie im Sinne eines offenen Katalogs kontinuierlich aktualisieren. Der Katalog bietet die Möglichkeit zu einzelnen Teilen der Beschreibung wie zur gesamten Handschrift Bilder beizugeben. Ebenso lassen sich ganze Handschriften einfügen, die dann seitenweise durchgeblättert werden können. NBM wird vollständig über das Internet verwaltet, von der primären Datenerfassung über die Revisionen bis zur abschließenden Veröffentlichung. Das System bietet Profile für verschiedene Benutzerrollen: Ein Koordinator richtet Benutzerkonten für die Katalogisierung ein, kontrolliert und begutachtet die Datensätze und überwacht den Inhalt des gesamten Webangebots. Katalogisierungsrichtlinien stellen ein Maximum an Einheitlichkeit bei den Beschreibungen sicher. Die Datenbank kann über einen OPAC auf der Webseite und über das Z39.50-Protokoll abgefragt werden. Bis heute hat NBM die Beschreibungen von mehr als 19.000 Handschriften veröffentlicht.

Abstract

Nuova Biblioteca Manoscritta (NBM) is an online catalogue of the manuscripts held in the libraries of the Veneto: approximately 90,000 items not limited by date or contents. Until now, this cultural heritage could only be represented in an incomplete manner by the printed catalogues which very often do not satisfy the demands of modern scholarship. The new project, funded by the Region of Veneto, overcomes this drawback. It started in 2003 and so far includes material from 38 libraries. NBM is an internet-based platform. Its nucleus is a central database to which the participating libraries add their data. Cataloguers use common authority files for names, titles, subjects, old shelfmarks, text types, textual genre and bibliographical information. The cataloguers have access to data already present, which they can update continuously, as of an open catalogue. It is possible to add images to every part of a descriptive record as well as to import a complete digitized manuscript, thus allowing for a page by page view. NBM is managed totally via internet, from first data input to revisions and final publication. The system provides several profiles corresponding to the roles in the project. A coordinator assigns user accounts for cataloguing, controls and reviews the records and supervises the content of the website. Guidelines for cataloguing ensure a maximum of conformity in the descriptions. The database is queried via the OPAC on the Website and via the Z39.50 protocol. NBM has so far described and published more than 18,000 manuscripts.

I manoscritti conservati nelle biblioteche del Veneto sono più di 90.000, senza contare gli innumerevoli carteggi o altra tipologia di materiale non in forma di codice. Questo patrimonio è attualmente accessibile per il tramite di cataloghi spesso antiquati, che di rado rispondono alle moderne esigenze degli studi, o di inventari manoscritti, consultabili solamente nella biblioteca che li conserva. Molte biblioteche, invece, non offrono alcuno strumento catalografico, di fatto impedendo di conoscere l'esistenza o la consistenza di questo o quel fondo di manoscritti. Per questo stato di cose, dal 2003 la Regione del Veneto ha deciso di avviare un progetto di catalogazione di tutti i manoscritti conservati nelle biblioteche venete, senza escludere dalla catalogazione particolari tipologie di manoscritto o stabilire limiti cronologici e di contenuto. La finalità era di mettere a disposizione della comunità scientifica e del pubblico più vasto tutto il patrimonio manoscritto attraverso un catalogo aperto, che avesse norme unitarie e condivise e permettesse di catalogare in modo rapido, con criteri scientificamente aggiornati e corretti, i manoscritti conservati nelle biblioteche della regione. Fu subito evidente che un progetto a cui aderivano biblioteche di diverse tipologie e numerosi catalogatori dislocati nel territorio avrebbe avuto la necessità di un importante lavoro di coordinamento e che compito fondamentale sarebbe stato quello del controllo e della validazione delle schede catalografiche prodotte. Fu creato un coordinamento scientifico che si appoggiò operativamente alla Biblioteca del Museo Correr di Venezia, la quale già aveva avviato un proprio progetto di catalogazione. Il coordinamento elaborò un modello di scheda di descrizione di tipo sommario, redatta secondo le norme previste dalla *Guida a una descrizione uniforme dei manoscritti e al loro censimento* (Roma 1990), che meglio rispondesse a una tipologia di materiale prevalentemente moderno e tenesse conto della quantità dei manoscritti, dei tempi della catalogazione e dei finanziamenti annuali disponibili. Elaborò inoltre delle linee guida per i catalogatori al fine di garantire una assoluta uniformità e omogeneità catalografica.

Come presentato da Eleuteri e Vanin (2005), Vanin e Eleuteri (2006 e 2007), Eleuteri (2007) e Vanin (2008), il progetto prese avvio nell'autunno del 2004, interessando 15 biblioteche (civiche, ecclesiastiche, private, di museo e di fondazioni), che presentarono ciascuna un progetto di catalogazione dei propri fondi manoscritti sprovvisti di catalogo a stampa. La catalogazione fu avviata con il software *Manus* dell'Istituto Centrale per il Catalogo Unico delle Biblioteche Italiane e per le Informazioni Bibliografiche (ICCU), cui le schede prodotte sarebbero state inviate per la loro successiva pubblicazione on line. *Manus* è un software che, installato localmente su PC, non consente una catalogazione partecipata, ma uno scarico periodico di dati in un'unica banca dati interrogabile on line.

Per la struttura del progetto *Manus*, emerse presto la considerazione che l'uniformità dei dati, fondamentale in un progetto di ampio respiro come quello veneto, era difficilmente conseguibile e che l'ICCU non avrebbe potuto garantire una rapida pubblicazione dei dati, per gli inevitabili problemi di controllo della validità scientifica delle schede, prodotte in maniera autonoma da diversi catalogatori. La Regione del Veneto decise allora di fare una scelta importante, significativa e nuova nell'ambito della catalogazione dei manoscritti: catalogare in maniera partecipata e gestire le catalogazioni direttamente sul web. Affidò al coordinamento scientifico la progettazione di NBM, uno strumento che consente una catalogazione partecipata utilizzando un'unica banca dati e si serve dei browser per immettere e pubblicare direttamente le descrizioni prodotte. In particolare, tutti i catalogatori condividono e incrementano le liste dei nomi, luoghi, titoli identificati, antiche segnature, argomenti e bibliografia. Lavorando su un'unica banca dati, creano e attingono a informazioni già strutturate, garantendo all'utente dati più uniformi, precisi e coerenti. Poiché la catalogazione avviene direttamente sul web, ogni informazione può essere corretta, modificata e integrata, nello spirito di un catalogo aperto in continuo aggiornamento. Attualmente, le biblioteche partecipanti al progetto sono 38; nei prossimi mesi aderiranno anche la Biblioteca nazionale Marciana di Venezia e la Biblioteca civica di Treviso. Dal 2008 il coordinamento scientifico del progetto è stato affidato all'Università Ca' Foscari Venezia, che ha elaborato la versione beta di NBM. Ad oggi i manoscritti catalogati e pubblicati sono più di 19.000, di cui circa 3.000 riversamenti di schede già prodotte con *Manus*.

Nuova Biblioteca Manoscritta è il sito web del progetto e il software di catalogazione. NBM è una applicazione web che integra tutte le funzioni necessarie alla catalogazione dei manoscritti, alla gestione degli utenti e alla pubblicazione dei contenuti. Per la parte informativa del sito web, NBM include un *Content Management System*, su cui è possibile inserire direttamente online le modifiche alle pagine web e i nuovi contenuti del «Diario» , un *blog* riguardante il progetto di catalogazione. L'area di catalogazione, cui si accede tramite login e password, raggruppa gli strumenti per i catalogatori, la messaggistica interna, i moduli di importazione e esportazione dati. Agli amministratori del sito sono riservate le funzioni per la gestione delle biblioteche, degli utenti, della biblioteca digitale, delle sezioni del sito.

Sia la parte pubblica che quella riservata si possono articolare in più sezioni, che corrispondono a diversi progetti di catalogazione facenti capo a diversi gruppi di catalogatori e amministratori. Le sezioni si differenziano per contenuti delle pagine web e hanno dedicate specifiche funzioni di ricerca sul catalogo. Questo vale anche per progetti che possono utilizzare alfabeti diversi in inserimento di dati e in ricerca da parte dell'utente. Nella parte pubblica sono presenti informazioni generali sul progetto e sul catalogo, sulle biblioteche partecipanti, la biblioteca digitale, una sezione di didattica del manoscritto, sussidi bibliografici e link utili per la catalogazione, la guida all'uso del software e le linee guida di catalogazione, contatti con i referenti. Dal sito web si potrà accedere anche ai moduli dinamici del catalogo, della biblioteca digitale, del diario, dell'accesso riservato. In presenza di più progetti, dalla pagina iniziale di NBM si accederà al catalogo generale e alle pagine statiche generali, ma anche ai siti e alle pagine dinamiche delle varie sezioni, con le dovute differenze e limitazioni. Un'area di ricerca è dedicata alla consultazione del catalogo in linea. La banca dati può essere interrogata mediante una ricerca semplice per parola/e secondo operatori booleani o mediante una ricerca avanzata, combinando cioè diversi campi della descrizione (la ricerca può essere effettuata anche per area geografica, per lingua, per bibliografia; sono anche ricercabili le parti a stampa presenti nei codici). Il risultato è una notizia breve che presenta numero progressivo, segnatura, datazione e un estratto della scheda che, se è corredata di immagini (alcune carte o una riproduzione integrale del manoscritto), avrà un simbolo linkabile accanto alla segnatura. Alla scheda possono essere legate immagini anche fornendo link a URL esterni. Dalla notizia breve si accede alla descrizione completa, che può essere stampata. Le eventuali immagini allegate sono corredate di una didascalia o commento. Sono presenti anche link al modulo di stampa pdf, txt e di richiesta riproduzioni (per quelle biblioteche che abbiano provveduto a fare l'upload dei moduli di richiesta riproduzioni). Nella parte inferiore della scheda sono presenti alcuni dati gestionali: data di pubblicazione, data della revisione del coordinatore, le forme varianti dei nomi indicizzati nella scheda, i legami del manoscritto con altri manoscritti (stessa legatura, stessa mano, stesso tipo di decorazione etc.), cui si può accedere per navigazione, soggetti, generi letterari, codici di contenuto, lingua, immagini. Dall'home page si accede alla «Biblioteca digitale» , che contiene i materiali digitalizzati (repertori utili alla catalogazione, bibliografia a stampa e non a stampa, manoscritti). La pagina organizzerà i materiali in maniera automatica per tipologia e secondo elenchi, con link alle pagine di consultazione vera e propria. La pagina «Diario» contiene una serie di notizie ordinate cronologicamente, su cui gli utenti possono lasciare commenti (blog del CMS). Sul sito della Regione del Veneto è attiva una newsletter. È a disposizione degli utenti l'archivio delle forme normalizzate dei nomi indicizzati nel catalogo.

Il progetto e la catalogazione, come abbiamo già accennato, sono gestiti totalmente in rete. Mediante login e password, assegnate sulla base di diversi profili, si accede all'area riservata di NBM. I profili previsti sono:

- catalogatore: accede esclusivamente ai manoscritti e alle biblioteche cui è associato;
- bibliotecario: visualizza i manoscritti della propria biblioteca ed è abilitato alle operazioni di esportazione dei dati in formato XML TEI-MS;
- revisore: corregge e approva le schede dei manoscritti, può intervenire nei moduli di gestione delle liste;
- coordinatore: esercita le funzioni di gestione degli utenti e delle biblioteche, impor-

tazione e esportazione dei dati in formato XML TEI-MS, importazione di elementi digitali, visualizzazione e modifica di tutte le schede di un determinato dominio;
- amministratore: ha accesso a tutti i moduli e a tutte le funzionalità, comprese quelle di gestione dell'applicazione, crea e gestisce i vari domini e il dominio principale di NBM.

Una tabella riassuntiva, che si aggiorna ad ogni variazione, consente all'amministratore di monitorare costantemente le catalogazioni di tutte le biblioteche partecipanti, nei diversi stati in cui si trovano. Nel corso del lavoro il catalogatore ed il revisore stabiliscono i diversi stati in cui ogni singolo manoscritto viene a trovarsi. Al momento della creazione della scheda e finché non è ritenuto pronto per l'approvazione del revisore, ogni manoscritto è in *lavorazione*; diventa *completato* quando il catalogare indica al revisore che il manoscritto può essere valutato per la pubblicazione; se il manoscritto è stato visionato, ma per via di errori o imprecisioni, non può essere pubblicato, viene considerato *da rivedere*; lo stato *corretto* segnala al revisore che le correzioni sono state apportate; infine, il manoscritto viene *pubblicato*. Ogni qual volta vi sia la necessità di apportare correzioni o modifiche, la scheda pubblicata tornerà allo stato in lavorazione, mentre per la ricerca resterà disponibile la scheda precedente, fino alla nuova pubblicazione. I passaggi nei diversi stati, tra la creazione della scheda e la sua pubblicazione, possono essere accompagnati da messaggi interni legati alla scheda fra revisore a catalogatore e sono caratterizzati dall'indicazione dell'autore e della data di revisione e risposta del catalogatore.

La scheda catalografica di NBM rispecchia i campi di *Manus*. Questo è stato stabilito fin dall'inizio per evitare ulteriori creazioni di software diversi, per restare strategicamente collegati ad un software nazionale e dunque anche per consentire un preciso import ed export dei dati. Ove possibile, si sono accorpati più campi in una stessa schermata, per rendere il lavoro di catalogazione meno frammentato. Inoltre, al di sotto dei moduli di inserimento è sempre visibile l'anteprima della scheda che si viene componendo nei diversi campi, così come apparirà all'utente. Anche questa soluzione è stata pensata per ovviare ad una delle difficoltà maggiori che si riscontrano nelle catalogazioni eseguite con uno strumento informatico, quella cioè di catalogare per campi slegati fra loro e di non avere mai sotto gli occhi la scheda nella sua interezza.

La catalogazione partecipata ha imposto la modifica di alcune pratiche di descrizione. Come abbiamo già detto, i catalogatori hanno in comune le liste di nomi, nomi nel titolo, luoghi, titoli, antiche segnature, tipologia del testo e del genere letterario, argomenti, bibliografia. La modifica di qualsiasi voce nelle liste comporta la modifica della voce in tutte le schede collegate. I nomi, formulati secondo le *Regole Italiane di Catalogazione per Autori* (RICA)[1], sono accompagnati, se identificati, dal rinvio al repertorio utilizzato

[1] Nel gennaio 2009 è stata pubblicata la bozza complessiva delle nuove Regole italiane di catalogazione (REICAT).

per l'identificazione e da altre informazioni che completano l'authority record (a seconda del tipo di nome, nazionalità, lingua, sesso, responsabilità, luogo e data di nascita e morte, indirizzo, note). Si possono creare rinvii tra la forma accettata del nome e le forme varianti o alternative nonché la bibliografia. Questo archivio è consultabile dagli utenti, che possono anche accedere ad un ulteriore archivio di nomi non identificati, ma per i quali è stata creata una forma normalizzata sulla base delle diverse forme presenti nei manoscritti e dell'uso; queste sono legate fra loro in una struttura ad albero, alla cui cima si trova la forma accettata, e da ognuna di esse si può accedere alle schede di manoscritti collegate. I nomi di luogo sono legati a stato (o parte di esso), regione, per poter consentire anche una ricerca per area geografica (luoghi di copia, di provenienza). I titoli identificati, cioè i titoli attribuiti sulla base di un repertorio o di un'edizione, contengono il nome dell'autore e altre eventuali responsabilità, unitamente al repertorio e/o all'edizione del testo utilizzati. Si possono creare legami fra titoli uniformi e forme varianti del titolo; sono preimpostati i legami tra le diverse forme di titolo (identificato, elaborato, presente, aggiunto), secondo una gerarchia fra titoli prestabilita. La citazione bibliografica completa è accompagnata dalla sua eventuale abbreviazione, che sarà utilizzata nelle schede; qui la bibliografia apparirà automaticamente in ordine cronologico. La lista delle antiche biblioteche/segnature intende favorire l'individuazione di provenienze omogenee dei manoscritti, con l'obiettivo di ricostruire virtualmente le biblioteche di appartenenza o anche di creare concordanze tra le attuali e le precedenti segnature. Nella descrizione interna è prevista l'indicazione della tipologia del testo e del genere letterario, secondo le forme di UNIMARC; in caso di opere anonime è necessario indicare anche l'argomento, recuperabile dalla specifica lista condivisa. Non si tratta dunque di una soggettazione vera e propria, quanto piuttosto di una o più indicazioni che facilitino l'individuazione del contenuto, secondo una pratica abbastanza diffusa nella catalogazione tradizionale dei manoscritti.

Di recente NBM è stata adeguata alla legge Stanca del 2004, che garantisce l'accessibilità ai mezzi informativi da parte di utenti disabili, ed è stata prescelta per far parte di CulturaItalia, il portale italiano della cultura, rispondendo ai requisiti previsti dal Ministero per i Beni e le Attività Culturali ed ottemperando ai principi di qualità suggeriti dal progetto europeo MINERVA. Il formato di scambio dei dati utilizzato è l'XML sviluppato dalla Text Encoding Initiative (TEI) per la descrizione dei manoscritti. La banca dati di NBM è interrogabile anche attraverso il protocollo Z39.50. Nel prossimo futuro altre biblioteche venete si aggregheranno al progetto, anche se non si esclude la partecipazione di biblioteche al di fuori della regione, che vogliano rendere i propri dati fruibili su NBM. Con gli ultimi sviluppi del software concernenti la molteplicità delle sezioni, cui abbiamo già accennato, sarà possibile anche gestire progetti tematici di catalogazione, mantenendo un unico data base, ma offrendo all'utente un'interfaccia di ricerca con parametri preimpostati. In particolare, assieme ad un gruppo di studiosi di università italiane si sta valutando la possibilità di utilizzare NBM per la realizzazione di un album paleografico digitale dei manoscritti greci conservati nelle biblioteche italiane, composto dal censimento dei dati identificativi essenziali dei singoli codici, da almeno una riproduzione fotografica per ciascuna unità codicologica e da una bibliografia il più possibile esaustiva ed aggiornata. Infine, in collaborazione con l'ICCU, che di recente ha predisposto la versione on line di *Manus* (*Manusonline*), in parecchi punti con soluzioni simili a quelle adottate da NBM, si sta valutando se costruire un OPAC centrale, che consenta di interrogare e ricercare contemporaneamente le due banche dati.

Bibliografia

CulturaItalia. <http://www.culturaitalia.it>.

Eleuteri, Paolo e Barbara Vanin. «Il catalogo on line dei manoscritti delle biblioteche del Veneto.» *Gazette du livre médiéval* 47 (2005): 31-38.

Eleuteri, Paolo. «La catalogazione in rete dei manoscritti delle biblioteche venete.» *Zenit e Nadir II. I manoscritti dell'area del Mediterraneo: la catalogazione come base della ricerca.* A cura di B. Cenni, C.M.F. Lalli e L. Magionami. Montepulciano: Thesan&Turan, 2007. 221-225.

Guida ad una descrizione uniforme dei manoscritti e al loro censimento. A cura di V. Jemolo e M. Morelli, Roma: ICCU, 1990.

Manus. Censimento dei manoscritti delle biblioteche italiane. Istituto Centrale per il Catalogo Unico delle Biblioteche Italiane e per le Informazioni Bibliografiche (ICCU). <http://manus.iccu.sbn.it>.

Manusonline. <http://193.206.221.40/manus>.

Nuova Biblioteca Manoscritta (NBM). <http://www.nuovabibliotecamanoscritta.it>.

Regole italiane di catalogazione (REICAT). 2009. <http://www.iccu.sbn.it/upload/documenti/REICA_bozza_complessiva_genn2009.pdf>.

TEI Consortium, ed. *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Chapter 10: Manuscript Description.* 1.3.0. Last updated on February 1st 2009. <http://www.tei-c.org/release/doc/tei-p5-doc/html/MS.html>.

UNIMARC Manual. 2nd ed. München, London: Saur, 1994.

Vanin, Barbara. «Nuova Biblioteca Manoscritta. Online Catalogue of Manuscripts Conserved in Libraries in the Veneto Region.» *Encyclopedia of Information Communication Technology* (ICT). A cura di Antonio Cartelli e Marco Palma. Hershey, Pennsylvania: IGI Global, 2008. 632-634.

Vanin, Barbara e Paolo Eleuteri. «Nuova Biblioteca Manoscritta. Catalogo in linea dei manoscritti delle biblioteche del Veneto.» *Bollettino dei Musei Civici Veneziani*, s. III 1 (2006): 113-117.

Vanin, Barbara e Paolo Eleuteri. «La Nuova Biblioteca Manoscritta della Regione del

Veneto.» *Conoscere il manoscritto: esperienze, progetti, problemi. Dieci anni del progetto Codex in Toscana.* A cura di M. Marchiaro e S. Zamponi. Firenze: Sismel, 2007. 145-152.

Z39.50. <http://www.loc.gov/z3950/agency>.

Il catalogo aperto dei manoscritti Malatestiani

Antonio Cartelli, Andrea Daltri, Paola Errani, Marco Palma, Paolo Zanfini

Riassunto

I manoscritti medievali conservati nella storica biblioteca cesenate sono 429, in parte notevole (343) collocati nella Biblioteca Malatestiana, fondata da Malatesta Novello, signore della città, alla metà del secolo XV. Ad essi si aggiungono codici della biblioteca privata di papa Pio VII (il cesenate Gregorio Barnaba Chiaramonti), otto corali commissionati dal cardinale Bessarione, sette corali di proprietà della Diocesi di Cesena e i manoscritti della Biblioteca Comunale o Comunitativa, costituitasi all'inizio dell'Ottocento con i fondi delle corporazioni religiose soppresse. In occasione del convegno di studi promosso nel 2003 per il 550° della fondazione, la Malatestiana ha presentato il *Catalogo aperto dei manoscritti Malatestiani*, nato dalla collaborazione con Antonio Cartelli e Marco Palma dell'Università di Cassino ma realizzato all'interno della Biblioteca. Esso è in sostanza un sistema informativo che fonda la sua struttura e le sue funzioni sull'utilizzo intensivo delle tecnologie dell'informazione e della comunicazione. Per la costruzione del catalogo aperto si è scelto di utilizzare l'applicativo WinISIS. Il sito del catalogo aperto è ospitato gratuitamente sul server pubblico della provincia di Forlì-Cesena. Il catalogo aperto contempla tre accessi in base alla lingua degli utenti: italiano, inglese e tedesco, ed è articolato in tre sezioni: la prima, contenente monografie e articoli, utili alla conoscenza della biblioteca e dei suoi fondi; la seconda presenta le descrizioni dei codici, la bibliografia relativa a ciascun manoscritto posseduto dalla biblioteca e le immagini che riproducono tutto o in parte le pagine dei manoscritti; l'ultima si basa su un sottosistema informativo ad accesso protetto e differenziato, molto simile ad un forum, in cui le persone interessate allo studio dei manoscritti della biblioteca possono pubblicare lavori oppure scambiarsi informazioni, formulare progetti e dibattere problemi di comune interesse.

Zusammenfassung

In der Historischen Bibliothek von Cesena befinden sich 429 mittelalterliche Handschriften, von denen ein Großteil (343) in der Biblioteca Malatestiana aufbewahrt wird. Diese ist in der Mitte des 15. Jahrhunderts von Malatesta Novella, seinerzeit Stadtherr zu Cesena, gegründet worden. Hinzu kommen Kodizes aus der Privatbibliothek Papst Pius' VII (dem Cesenaten Gregorio Barnaba Chiaramonti), acht Choralbücher des Kardinals Bessarione, sieben Choralbücher aus dem Besitz der Diözese Cesena und die

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 13–23.

Handschriften aus der Kommunalbibliothek, deren Archivbestände aus den Bibliotheken der zum Beginn des 18. Jahrhunderts aufgelösten kirchlichen Orden stammen. Im Jahre 2003 präsentierte die Bibliothek auf einer Tagung anlässlich ihres 550jährigen Bestehens den *Offenen Handschriftenkatalog der Biblioteca Malatestiana* als Ergebnis eines Projekts von Antonio Cartelli und Marco Palma von der Universität zu Cassino, das aber in der Bibliothek selbst umgesetzt wurde. Es handelt sich hierbei im Wesentlichen um ein Informationssystem, dessen Aufbau und Funktionen sich auf die intensive Verwendung von Informations- und Kommunikationstechnologien stützen. Für die Erarbeitung des Offenen Katalogs wurde WinISIS eingesetzt. Die Webseite des Katalogs wird kostenfrei von dem öffentlichen Server der Provinz Forlì-Cesena gehostet. Das Katalogangebot steht dem Benutzer auf Italienisch, Englisch und Deutsch zur Verfügung und ist in drei Sektionen gegliedert: die erste Sektion umfasst Monographien und Artikel über die Bibliothek und ihre Bestände; die zweite Sektion bietet Handschriftenbeschreibungen und Bibliographien zu jeder einzelnen Handschrift der Bibliothek sowie Abbildungen der Handschriften als Ganze oder in Teilen; in Gestalt eines Forums ist die dritte Sektion interessierten Nutzern nur nach vorangehender Anmeldung zugänglich und der Veröffentlichung von Forschungsergebnissen, dem Austausch von Informationen, der Ausarbeitung neuer Projekte und der Diskussion gewidmet.

Abstract

In the Malatestiana Library, built in the mid-15th century by Malatesta Novello, Lord of Cesena, 343 manuscripts are housed. It also houses two 15th century local liturgical series: seven choral books from the Cathedral and eight from the Franciscan convent. 59 manuscripts dating from the 12th to the 15th century belong to the Piana Library, the private collection of Pope Pius VII (Barnaba Chiaramonti). Twelve more manuscripts belonged to the town library, which was formed at the beginning of the 19th century with the books once owned by the dissolved religious houses. The total number of manuscripts is 429. In 2003, on the occasion of the 550th anniversary of the foundation of the Library, the staff of the Library supported the idea of an *Open Catalogue of the Malatestiana Manuscripts*, which makes intense use of information and communication technology and makes available all the documentation on the Net, retrieving and updating previous and recent information. Conceived by Antonio Cartelli and Marco Palma of the University of Cassino and realized by the Malatestiana Library, the Open Catalogue offers texts on the Library and its manuscript collections, descriptions of manuscripts (that is, previous printed catalogues and new descriptions especially produced or commissioned by the Malatestiana Library), a bibliography constantly updated and a rich section of images of the manuscripts. By filling in a form available in the website, it is possible to enter the Forum, where scholars or persons interested in the library's manuscripts can receive information, contribute with their observations, as well as publish their studies on the Malatestiana manuscripts. The Open Catalogue of Manuscripts in the Malatestiana Library can be accessed starting from the Province of Forlì-Cesena portal, with direct access from the new Library web page. The database was autonomously constructed with a WinISIS application.

Il catalogo aperto dei manoscritti della Biblioteca Malatestiana, presentato a Cesena per il convegno di studi «Il dono di Malatesta Novello» nel marzo 2003 (Cartelli et al. 2006), si basa su un'idea proposta in occasione di un incontro di ‹filosofi della rete› a Cork, in Irlanda, nel giugno 2002 (Cartelli e Palma 2002).

Il catalogo aperto è stato pensato per essere utilizzato da ogni biblioteca in possesso di fondi manoscritti con l'obiettivo di restituirle il ruolo centrale di produzione culturale che essa aveva nei secoli passati. Si contraddistingue inoltre per la sua elasticità e dinamicità nei confronti delle corrispondenti chiusura e staticità del catalogo a stampa. Esso è, per molti versi, un sistema informativo (nel senso più propriamente informatico del termine), che consta dell'insieme delle risorse umane, hardware e software necessarie a gestire informazioni documentarie e fonda la sua struttura e le sue funzioni sull'utilizzo intensivo delle tecnologie dell'informazione e della comunicazione.

Il catalogo aperto, nella sua struttura generale, è articolato in diverse sezioni, da intendere in maniera flessibile almeno nella fase di avvio, nel senso che la biblioteca che decide di adottarlo ne può attivare di meno o di più, a seconda della sua disponibilità e capacità, ed in maniera tale che ciascuna di esse possa essere gestita nei tempi e nei modi che le sue risorse umane e finanziarie consentono:

1. La prima sezione è destinata a contenere materiali, già editi o prodotti per l'occasione o che possono essere previsti in futuro, utili alla conoscenza della biblioteca e dei suoi fondi, come monografie e articoli. Questi documenti vogliono offrire all'utente che si avvicina alla biblioteca un quadro coerente dell'insieme dei materiali di cui fanno parte gli esemplari che interessano o che sono oggetto di studio.
2. Nella seconda sezione è prevista la bibliografia dei manoscritti in possesso della biblioteca, articolata per segnatura, in ordine alfabetico o cronologico.
3. La terza sezione presenta le precedenti descrizioni a stampa dei codici o anche, opportunamente digitalizzate, quelle contenute negli antichi inventari manoscritti. Ovviamente vi devono figurare anche le nuove descrizioni, secondo standard definiti ma non tali da impedire ogni forma di libertà ai redattori.
4. Nella quarta sezione trovano posto le immagini che riproducono in tutto o in parte le pagine dei manoscritti, per le quali, in generale, possono essere adottate diverse soluzioni tecniche. Ciò che si intende proporre con il catalogo aperto è un caso tipico di compromesso: immagini non dettagliatissime, il cui download risulterebbe altrimenti assai pesante, ma di una risoluzione e con caratteristiche di luminosi-

tà e contrasto tali da garantire la loro intelligibilità, e, soprattutto, in numero tale da documentare il massimo numero possibile, potenzialmente tutti, i codici della biblioteca.

5. L'ultima sezione rappresenta una novità rispetto al normale utilizzo della rete in campo paleografico. Essa si basa su un sottosistema informativo ad accesso protetto e differenziato, molto simile ad un forum o ad una chat, in cui le persone interessate allo studio dei fondi manoscritti della biblioteca possono pubblicare, con tutte le garanzie relative alla privacy e alla protezione dei diritti d'autore, i lavori concernenti i manoscritti della biblioteca stessa, oppure possono scambiarsi informazioni, formulare progetti e dibattere problemi di comune interesse.

Si potrebbe pensare, a questo punto, che il catalogo aperto dei manoscritti Malatestiani, in quanto concretizzazione di un progetto, sia quella che, in gergo prettamente informatico, viene chiamata una implementazione del progetto stesso. Occorre però sgombrare il campo da equivoci e chiarire che si tratta di una cosa profondamente diversa.

Infatti le linee guida del catalogo aperto non prevedono la definizione di standard relativi alla risoluzione delle immagini, alla modalità della loro visualizzazione, all'utilizzo di software specifici per la creazione e la gestione di basi dati, cioè, in altre parole, non si propongono di definire nel dettaglio il processo di realizzazione del sistema informativo; si ritiene opportuno, infatti, che ogni biblioteca che voglia adottare il progetto e realizzarlo utilizzi i mezzi e le tecniche per le quali ha delle competenze consolidate, in grado di garantire non solo l'avvio dell'iniziativa, ma anche il suo mantenimento e la sua eventuale evoluzione. Inoltre il catalogo aperto, pur prevedendo cinque sezioni di base, non le esaurisce, nel senso che è pensabile che esigenze specifiche di una realtà locale possano richiedere una diversa articolazione delle sezioni previste dal progetto.

I manoscritti medievali conservati nella storica biblioteca cesenate sono 429, in parte notevole (126) prodotti nell'ambito del progetto culturale di Malatesta Novello, che costruì la Malatestiana alla metà del secolo XV. Ad essi si aggiungono esemplari di diversa provenienza, tra i quali si annoverano l'antico fondo conventuale francescano, la collezione del medico umanista riminese Giovanni di Marco, alcuni codici della biblioteca privata di papa Pio VII (il cesenate Barnaba Chiaramonti), otto corali fatti produrre dal cardinal Bessarione, sette corali di proprietà della diocesi di Cesena e i manoscritti della Biblioteca Comunale o Comunitativa, costituitasi all'inizio dell'Ottocento con i fondi delle corporazioni religiose soppresse.

Per la costruzione del catalogo aperto dei manoscritti Malatestiani si è scelto di utilizzare l'applicativo Winisis (CDS/ISIS). I motivi di questa scelta, rispetto all'impiego di software più noti e di maggiore diffusione, sono molteplici: la gratuità del prodotto, la sua distribuzione da parte dell'UNESCO, che ne assicura il mantenimento e il costante aggiornamento nel tempo, il patrimonio di conoscenze già presente all'interno della Malatestiana, che ha consentito di evitare l'oneroso ricorso a ditte esterne per l'elaborazione e la gestione del progetto informatico, e infine la disponibilità di un programma, denominato Wxis, che permette d'interfacciare i database di Winisis con un server web.

L'evidente e predominante motivazione economica trova una giustificazione anche nella volontà di mantenere un controllo diretto di tutti gli aspetti del progetto e di rifiutare in modo programmatico che il prodotto informatico possa trovare un approdo definitivo. Pertanto, in accordo con la filosofia che sottende la realizzazione del catalogo aperto, anche la sua struttura informatica si configura come aperta e in continua evoluzione: ripensamenti, correzioni, implementazioni di nuove funzionalità hanno infatti ritmato il tempo trascorso dalla pubblicazione in rete della prima versione.

La struttura del catalogo aperto è attualmente costituita da otto banche dati:

- Manoscritti
- Bibliografia
- Citazioni
- Icone dei manoscritti
- Immagini dei manoscritti
- Immagini dei testi
- Iscritti al Forum
- Messaggi inviati al Forum

I database non sono inseriti all'interno di un reticolo relazionale in quanto Winisis è in grado di gestire soltanto relazioni «uno a uno». Per aggirare questo limite è stata adottata una struttura che consente di correlare i diversi database soltanto sul web mediante il valore attribuito a determinati campi che fungono da legame e attivano specifici input di ricerca. Sotto questo profilo la struttura informatica è specificatamente orientata alla pubblicazione in rete.

La figura 1 mostra i campi che attivano i legami tra i diversi database.

Il sito del catalogo aperto, ospitato gratuitamente sul server pubblico della provincia di Forlì-Cesena, contempla tre accessi in base alla lingua degli utenti: italiano, inglese e tedesco.

Il sito è costituito da una parte statica in linguaggio HTML, sostanzialmente limitata alle pagine che introducono le diverse sezioni e ai testi non digitalizzati, e da una preponderante parte dinamica che mediante form CGI e il linguaggio di scripting di Wxis consente d'interagire con i database di Winisis.

Il sito è articolato in quattro sezioni:

- Progetto
- Testi
- Manoscritti
- Forum

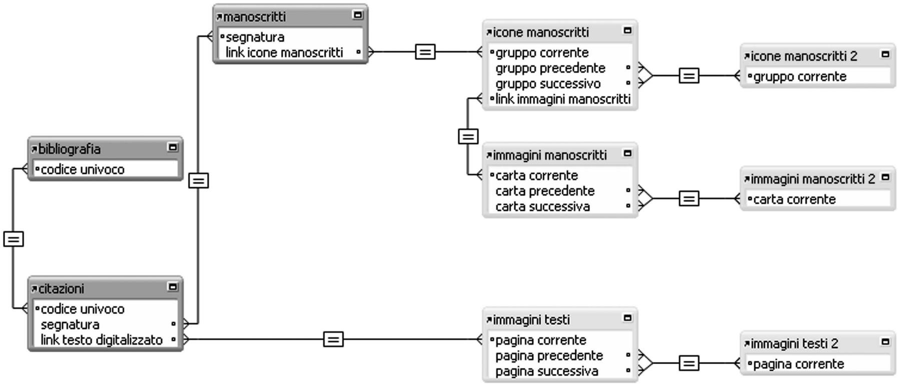


Figura 1. Campi che attivano i legami tra i diversi database.

Nella prima sezione si forniscono le coordinate culturali e metodologiche del progetto. Tra le altre sezioni si è tentato di assicurare la massima integrazione e navigabilità possibile.

Nella sezione Testi, organizzata al proprio interno per aree tematiche e soggetta a un costante incremento, sono resi disponibili saggi, articoli e fonti documentarie sulla Biblioteca Malatestiana e il suo patrimonio manoscritto (attualmente 128). Le aree tematiche sono le seguenti:

- La Biblioteca Malatestiana (in generale)
- La struttura edilizia e la storia architettonica
- La cultura umanistica cesenate
- Gli antichi inventari e i cataloghi storici
- Lo *scriptorium* cesenate e i copisti
- Gli aspetti codicologici e filologici
- La miniatura e l'iconografia
- I manoscritti ebraici e greci
- Le collezioni (Giovanni di Marco, Piana, Corali)
- Il catalogo aperto

Nella scelta dei testi sono stati privilegiati i contributi considerati ormai «classici» e quelli di più difficile reperibilità. La consultazione può avvenire in formato testo (HTML) o in formato digitale (formato JPEG di 550/700 pixel di ampiezza). In seguito sono state inaugurate due sottosezioni. La prima contiene un elenco delle tesi di laurea dedicate ai manoscritti Malatestiani, alcune delle quali già liberamente consultabili. La seconda, al momento limitata a poche fonti, intende offrire una raccolta delle descrizioni della Malatestiana tratteggiate dai viaggiatori italiani e stranieri che l'hanno visitata nel corso del tempo.

La sezione Manoscritti rappresenta la chiave d'accesso alle informazioni memorizzate nei database di Winisis. La sezione, anch'essa soggetta a un costante accrescimento, incarna la prima accezione del catalogo aperto, ovvero un contenitore in continua evoluzione che offre un'immagine dinamica dello stato delle conoscenze sui manoscritti Malatestiani. Attualmente sono disponibili 4631 voci bibliografiche e 1240 descrizioni (204 delle quali appositamente redatte nell'ambito del progetto secondo le regole dello standard nazionale *Manus*; cf. Guida 1990). I manoscritti integralmente digitalizzati sono 97 per un totale di 28197 scatti. Le immagini dei manoscritti sono state acquisite con fotocamera digitale con una risoluzione di 72 DPI. Per la pubblicazione sul web viene utilizzato il formato JPEG con un ampiezza pari a 700/1400 pixel.

Sono previste cinque modalità diverse d'interrogazione tra loro integrate:

- Ricerca per segnatura
- Ricerca semplice
- Ricerca avanzata
- Ricerca per liste
- Ricerca per immagini

La ricerca per segnatura consente d'individuare i record relativi a ogni manoscritto partendo dalla sua segnatura. La maschera d'interrogazione consente di attivare tre tipi di ricerca, recuperando rispettivamente la bibliografia, le descrizioni e le immagini disponibili per il manoscritto desiderato. I risultati sono visualizzati in tre pagine distinte, ma tra loro navigabili mediante appositi pulsanti. Nel caso di una ricerca relativa alla bibliografia o alle descrizioni i risultati sono ordinati in base al criterio prescelto (alfabetico o cronologico). In calce a ogni voce bibliografica è riportato, qualora sia stato redatto, l'abstract; mentre a fianco, se si tratta di una descrizione o se è consultabile la versione integrale del contributo, compare un link che consente di visualizzarne il testo in una finestra indipendente. Qualora sia stata effettuata una ricerca per la tipologia immagini, la pagina dei risultati elenca quelle disponibili suddivise in gruppi di 24 carte ciascuno. Cliccando su un'icona si apre una finestra nella quale viene visualizzata l'immagine prescelta. Rimanendo posizionati sulla finestra è possibile navigare all'interno del manoscritto utilizzando gli appositi pulsanti per richiamare le immagini delle carte precedenti e successive.

La ricerca semplice e la ricerca avanzata permettono d'impostare una strategia più articolata e complessa. La prima maschera d'interrogazione consente di formulare un'espressione di ricerca in relazione a due ambiti distinti, l'area del manoscritto (autore, titolo, copista, data topica e data cronica) e quella dei contributi critici (autore, titolo, rivista, data e parole dell'abstract), sia in riferimento a tutti i campi di ciascuna area, sia all'interno di un campo specifico. Nella seconda maschera d'interrogazione è invece possibile inserire un'espressione di ricerca in corrispondenza di ognuno dei campi che danno accesso alle informazioni contenute nel database. Nell'area del manoscritto, oltre a quelli della maschera di ricerca semplice, anche il supporto, la scrittura, le parole della descrizione e una serie di filtri (composito, palinsesto, exemplar, manoscritto peciato). Il ricorso agli operatori logici booleani e la possibilità di effettuare l'interrogazione per termini esatti o troncati conferiscono ulteriori opportunità di combinare e specificare meglio la formulazione della propria espressione di ricerca. Inoltre, cliccando sull'apposito bottone posizionato all'altezza di ciascun campo, è possibile accedere al relativo authority file e importare la voce desiderata. I risultati dell'interrogazione sono visualizzati nella pagina di risposta corrispondente alla tipologia di ricerca prescelta (bibliografia, descrizioni o immagini) secondo l'ordinamento desiderato (alfabetico, cronologico o per segnatura).

La ricerca per liste consente di effettuare un'interrogazione posizionandosi all'interno degli authority file del database. Nella maschera iniziale, dopo avere selezionato un campo attinente a una delle due aree sopra menzionate, è possibile sia digitare la voce desiderata, sia lasciare vuoto il campo d'immissione. In quest'ultimo caso nella maschera successiva viene visualizzato l'intero elenco dei termini indicizzati; nel primo caso invece la visualizzazione inizia dalla voce prescelta o, in assenza di questa, da quella immediatamente successiva in ordine alfabetico. Operando la selezione desiderata è possibile lanciare un'interrogazione che utilizzando l'operatore logico «or» di default consente di visualizzare i risultati in modo analogo alle altre modalità di ricerca.

La ricerca per immagini permette di ottenere l'elenco dei manoscritti integralmente digitalizzati. I risultati dell'interrogazione sono visualizzati in ordine di segnatura.

La figura 2 riassume le diverse modalità di ricerca, illustrandone i percorsi rispettivi e i database coinvolti.

Nell'ottobre 2006 è stata inaugurata la sezione Miniature, che consente di eseguire una ricerca – per miniatore, secolo di esecuzione e soggetto rappresentato – e di visualizzare le immagini relative sul sito dell'Istituto internazionale di storia economica «F. Datini» di Prato in base a una convenzione stipulata con la Biblioteca Malatestiana.

Nel novembre 2008 è stata inaugurata la sezione Peciae, che fornisce un elenco di tutti i manoscritti che contengono indicazioni di pecia con la possibilità di visualizzare le immagini relative.

La sezione Forum costituisce l'ambito nel quale trova attuazione uno dei principi base del catalogo aperto, ovvero uno spazio offerto alla collaborazione dell'utenza per segnalare materiali, scambiare informazioni, pubblicare contributi inediti. La possibilità di accedere a questa sezione è subordinata alla registrazione dell'utente. All'interno dell'area riservata l'iscritto può prendere visione dei contributi pubblicati nella sezione Testi, consultare la lista degli iscritti, inviare messaggi compilando l'apposito form, effettuare ricerche nell'archivio di quelli spediti. Attualmente il Forum conta 285 iscritti.

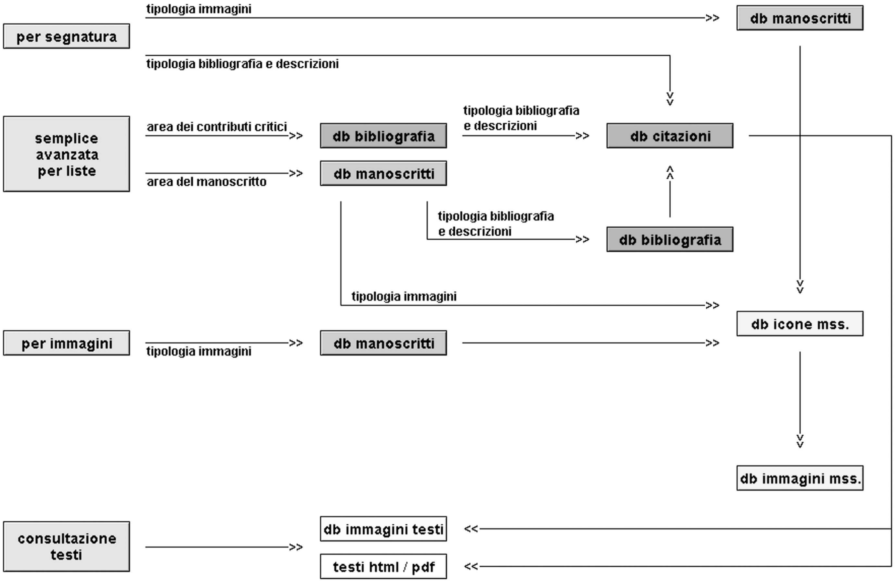


Figura 2. Modalità di ricerca con i percorsi rispettivi e i database coinvolti.

Nel corso del tempo la sezione si è arricchita di altre due funzionalità: la bibliografia partecipata e i cantieri aperti.

La prima, inaugurata all'inizio del 2005, consente a tutti gli iscritti al Forum di cooperare alla costruzione della bibliografia sui manoscritti Malatestiani creando direttamente in rete, mediante una procedura guidata, un nuovo record bibliografico con le relative citazioni. La visibilità del record è condizionata dalla validazione dell'amministratore del sito, che espletando la sua funzione di controllo può intervenire per modificarlo (nel caso sia stato commesso qualche errore) o cancellarlo (qualora rappresenti il duplicato di una notizia già esistente).

I cantieri aperti sono uno spazio riservato a tutti coloro che desiderano contribuire attivamente allo studio di alcuni dei più antichi manoscritti della Biblioteca Malatestiana. Sono stati finora inaugurati i cantieri sull'Isidoro Malatestiano (S.XXI.5) del IX secolo e sull'Evangeliario della Piana (3.210), datato 1104.

Nell'ambito del primo cantiere sono stati prodotti una descrizione esterna del codice e una serie di contributi su vari aspetti dell'esemplare. Inoltre è stato approntato un confronto visivo con l'apografo diretto del codice cesenate, il Marciano lat. II 46, di circa tre secoli più tardo. Con la collaborazione di due giovani studiosi, Anna Bellettini e Filippo Ronconi, è in costruzione una poligrafia che sarà intitolata Biografia di un manoscritto, prevista in uscita per il 2009 presso la casa editrice Viella di Roma La pubblicazione del volume comporterà ovviamente la chiusura del cantiere.

Nel secondo cantiere figurano attualmente una descrizione esterna e un'analisi grafica dell'Evangeliario, un manoscritto in splendida romanesca finora sostanzialmente ignoto alla letteratura paleografica.

L'interazione con l'utenza non si limita alla sezione Forum, ma prevede anche l'invio, in concomitanza con l'immissione nel sito di nuovi materiali, di una newsletter a tutti gli iscritti (giunta al numero 50). Le stesse informazioni sono messe a disposizione anche di tutti i visitatori nella pagina delle News. Nel dicembre 2004, al duplice scopo di intercettare i desiderata degli utenti e di orientare i futuri sviluppi del catalogo aperto, è stato inoltre promosso sul sito un sondaggio.

Nel corso dell'ultimo anno le visite al sito, al netto di robot e spider, hanno mantenuto una media giornaliera superiore alle 100 unità. Questi dati sono pubblicati periodicamente nella sezione Rapporti statistici.

In conclusione ci sembra opportuno sottolineare che si è in larga misura realizzato quanto ci si augurava nella presentazione del progetto, in particolare:

- l'attribuzione della funzione di coordinamento ai bibliotecari istituzionali, che hanno provveduto anche a produrre in proprio materiali (descrizioni e bibliografia);
- il costante contatto della biblioteca con le istituzioni e gli studiosi interessati alla ricerca sui manoscritti Malatestiani;
- l'utilizzazione di personale esterno giovane e qualificato, che può mettere a frutto la formazione ricevuta negli studi universitari.

In linea generale è importante l'effetto che esso potrebbe avere sul modo di effettuare studi e ricerche nel campo dei manoscritti. Asserire che la rete abbia contribuito a ridurre l'isolamento dello studioso, consentendo la rapida condivisione delle sue idee e dei risultati delle sue ricerche, o che grazie ad essa è divenuto possibile l'accesso da parte di un maggior numero di persone a informazioni e documenti prima patrimonio di pochi, è cosa ben nota e senz'altro vera, ma è anche sicuramente riduttiva, perché non rende giustizia ai profondi cambiamenti che le nuove tecnologie dell'informazione e della comunicazione hanno prodotto e possono ancora produrre nella società contemporanea e in particolare nella comunità scientifica. Mediante le tecnologie che essa rende disponibili è divenuto infatti possibile favorire la condivisione di esperienze di apprendimento e costruzione comune della conoscenza, in quanto consente ai soggetti che lo utilizzano di non limitarsi alla pura e semplice fruizione delle informazioni in esso contenute, ma piuttosto di identificarsi in un impegno comune, di praticare un'attività ed un linguaggio specifici e di possedere un repertorio condiviso di strumenti di lavoro. L'infrastruttura comunicativa caratteristica dell'ultima sezione del catalogo aperto (il forum) è tra i principali strumenti del processo appena descritto, poiché consente ai soggetti interessati di superare i limiti spazio-temporali del contesto fisico nel quale essi condividono le loro esperienze.

Lo sforzo futuro dovrà essere indirizzato al miglioramento di contenuti e aspetti formali, ma soprattutto all'ulteriore dimostrazione della versatilità del progetto, dove le potenzialità di implementazione delle modalità di ricerca e di aree dedicate o tematiche evidenzino le soggettive peculiarità dei fondi, marcando la differenza che lo separa dalle cumulative e più generali banche dati. D'altro canto, come per tutti i progetti simili a stretta connotazione identitaria, si corre il rischio di creare eccellenze qualitativamente molto valide, ma che rimangono tuttavia non comunicanti nell'eterogeneo *mare magnum* delle basi dati tematiche. L'utopica ambizione di giungere in futuro alla nascita di un portale di ricerca su un'ampia lista delle esperienze esistenti è inevitabilmente l'unica soluzione per acquisire una maggiore visibilità dei singoli progetti ed un incremento di fruizione e partecipazione da parte della comunità degli studiosi.

Bibliografia

Bellettini, Anna, Paola Errani, Marco Palma e Filippo Ronconi. *Biografia di un manoscritto. L'Isidoro Malatestiano S.XXI.5.* Con il contributo di Antonella Cesarini, Gaetano Martini, Anna Nardo e Nicola Tangari. Roma: Viella, 2009.

Cartelli, Antonio e Marco Palma. «Towards the Project of an Open Catalogue of Manuscripts.» *Proceedings of the Informing Science + Education Conference* (Cork, 19-21 June 2002). <http://proceedings.informingscience.org/IS2002Proceedings/papers/Carte188Towar.pdf>.

Cartelli, Antonio, Andrea Daltri, Marco Palma e Paolo Zanfini. «Il catalogo aperto dei manoscritti della Biblioteca Malatestiana: un primo bilancio.» *Il dono di Malatesta Novello.* Atti del convegno (Cesena, 21-23 marzo 2003). A cura di Loretta Righetti e Daniela Savoia. Cesena: Il ponte vecchio, 2006. 493-501.

Catalogo aperto dei manoscritti della Biblioteca Malatestiana. <http://www.malatestiana.it/manoscritti>.

CDS/ISIS database software. UNESCO and Information processing tools. <http://portal.unesco.org/ci/en/ev.php-URL_ID=2071>.

Guida a una descrizione uniforme dei manoscritti e al loro censimento. A cura di Viviana Jemolo e Mirella Morelli. Roma: ICCU, 1990.

Manus. Censimento dei manoscritti delle biblioteche italiane. Istituto Centrale per il Catalogo Unico delle Biblioteche Italiane e per le Informazioni Bibliografiche (ICCU). <http://manus.iccu.sbn.it>.

Die Sammlung Georg Rörers (1492–1557) Ein interdisziplinäres und multimediales Erschließungsprojekt an der Thüringer Universitäts- und Landesbibliothek Jena

Christian Speer

Zusammenfassung

Thema dieses Beitrags ist das interdisziplinäre und multimediale Forschungsprojekt an der Thüringer Universitäts- und Landesbibliothek Jena, welches die Sammlung von Handschriften und Drucken des Luthermitarbeiters Georg Rörer (1492–1557) erschließen und auf einer Online-Plattform (University Multimedia Electronic Library of Jena) präsentieren wird. Wichtigster Inhalt dieser Datenbank werden die Digitalisate und entsprechenden Katalogisate der Handschriften und Drucke aus Rörers Nachlass sein, wobei die auf mehrere Institutionen verteilten Handschriften in einer virtuellen Bibliothek erstmals wieder zusammengeführt werden. Um die wissenschaftliche Auseinandersetzung mit der Rörer-Sammlung zu fördern, werden die Katalogisate zu den Handschriften und Drucken bewusst als *work in progress* präsentiert und die potentiellen Nutzer zur Mitarbeit und Verbesserung eingeladen. Dafür werden nicht nur umfangreiche Recherchewerkzeuge angeboten, sondern auch die jeweilige Quellenüberlieferung, Erschließungsmittel, Editionsstand und Forschungsliteratur übersichtlich präsentiert. Dies ermöglicht jedem interessierten Wissenschaftler, in den Katalogisaten zu recherchieren, sofort alle bisher nicht gedruckten Texte zu identifizieren, schwierige Lesarten zu überprüfen, nach besonderen Merkmalen zu suchen oder die bisher nicht identifizierten Texte einer eigenen Untersuchung zu unterziehen. Erst diese als Gesamtheit präsentierten Handschriften und Drucke und die dazugehörigen Katalogisate und Hilfsmittel werden es ermöglichen, die Rezeption der präsentierten Texte, deren Reproduktion, Interpretation und/oder Umwandlung als Einheit, beispielsweise im Sinne einer »Ideengeschichte«, zu erforschen. Letztlich ist es ein Anliegen des Projekts, die Rolle und Funktion Rörers, respektive seiner Sammlung, im Kontext der Reformation neu zu bewerten und der Rezeptionsforschung sowie der Reformationsforschung neue Impulse zu verleihen.

Abstract

The topic of this contribution is to describe an interdisciplinary and multimedia research project at the Thuringian University and Federal State Library Jena, which will

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 25–34.

present a collection of manuscripts and printings of the Luther co-worker Georg Rörer (1492–1557) in an online information system (University Multimedia Electronic Library of Jena). The most important part of this database will be the digitized and cataloged manuscripts and prints from Rörer's Nachlass, whereby the manuscripts, scattered among several institutions, will be brought together for the first time in a virtual library. In order to promote the scholarly debate with the Rörer collection, the catalogue data sets belonging to the manuscripts and prints will be presented as work in progress, and potential users are invited to engage with and improve the data. Therefore, extensive search tools are offered along with the respective original sources, research tools, overviews to critical text editions and research literature. This allows for interested scholars to investigate the catalogues, to immediately identify unpublished texts, to examine difficult readings, to search for special characteristics or to submit to investigation into text as yet not classified. It will be possible to study as an entity the reception of the presented texts, their reproduction, interpretation and/or transformation, for instance in the meaning of an "Ideengeschichte". In the long the project allows to evaluate the role and importance of Rörer and his collection and to give new impetus to reformation studies.

1 Einleitung

Spätestens seit 2007, als eine Notiz des Luthermitarbeiters Georg Rörer zum Thesenanschlag Martin Luthers in einem Jenaer Druck wiederentdeckt wurde (vgl. den Sammelband Ott/Treu), stehen die Handschriften, Inkunabeln und frühen Drucke der Thüringer Universitäts- und Landesbibliothek Jena (ThULB) im Fokus der Wissenschaft. Ein besonderes Augenmerk richtet sich seitdem auf die Sammlung Georg Rörers, die für die Erforschung der Reformationsgeschichte einen außerordentlichen Stellenwert besitzt. Die Bedeutung jener Kollektion von 35 Handschriften (siehe das Beispiel in Abb. 1) und drei Drucken besteht vor allem darin, dass sie für die Überlieferung der Werke Martin Luthers, aber auch der Schriften Johannes Bugenhagens, Philipp Melanchthons und weiterer Reformatoren eine zentrale Textbasis darstellt. Von den Herausgebern der kritischen Lutherausgabe (WA) sind die Mit- bzw. Abschriften Rörers von Predigten, Vorlesungen, Tischreden etc. Martin Luthers zwar zum Druck gebracht worden, aber alles nicht-lutherische blieb zum großen Teil von der systematischen Erschließung ausgenommen. Selbst die Texte eines so bedeutenden Reformators wie Johannes Bugenhagen harren, von Ausnahmen abgesehen, ihrer kritischen Edition (vgl. den Forschungsstand zu Bugenhagen in Lorentzen). Diese einseitige Rezeption und damit nur selektive Erschließung der Handschriften Rörers ist vor allem einer auf Luther zentrierten Forschung der Vergangenheit geschuldet, die hier nicht weiter dargelegt werden soll. Vielmehr wird im Folgenden ein Projekt vorgestellt, das diese Forschungslücke schließen

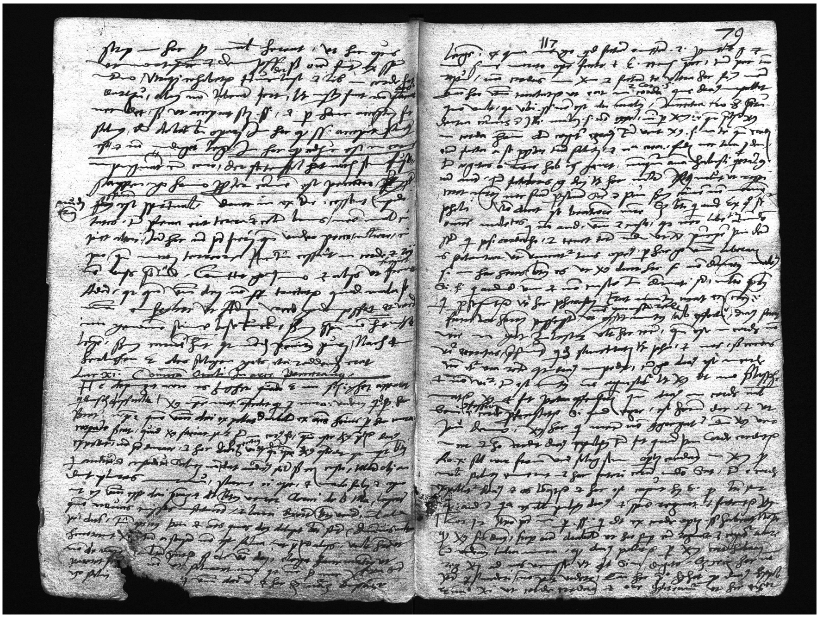


Abbildung 1. ThULB Jena, Ms. Bos. o. 17c, fol. 78v-79r, Predigt Johannes Bugenhagens, Schreiber: Georg Rörer (17 x 22 cm).

möchte. Zuvor müssen jedoch einige Sätze dem Leben und Werk des Georg Rörer gewidmet werden, um die Besonderheiten seiner Sammlung zu verstehen.

2 Leben und Werk des Georg Rörer (1492–1557)

Georg Rörer wurde am 1. Oktober 1492 in Deggendorf (bei Passau) geboren (vgl. zu G. Rörer Ott und die dort angegebene ältere Literatur). 1520 erwarb er den Grad eines Magisters an der Leipziger Universität. Seit 1522 ist er in Wittenberg nachweisbar. Dort studierte er weiter, eine Promotion ist aber nicht belegt. 1525 wurde er durch Martin Luther († 1546) zum ersten evangelischen Diakon der Stadtpfarrkirche in Wittenberg ordiniert. Im selben Jahr heiratete er Johanna Bugenhagen († 1527), die Schwester des Reformators Johannes Bugenhagen († 1558), mit der er im Hause Luthers wohnte. 1526 taufte er Martin Luthers Sohn Johannes. Nicht zuletzt durch diese persönliche Nähe wurde Rörer einer der engsten Vertrauten Luthers. So war es ihm aus unmittelbarer Kenntnis oder aus Mitschriften Dritter möglich, zahlreiche Predigten, Vorlesungen,

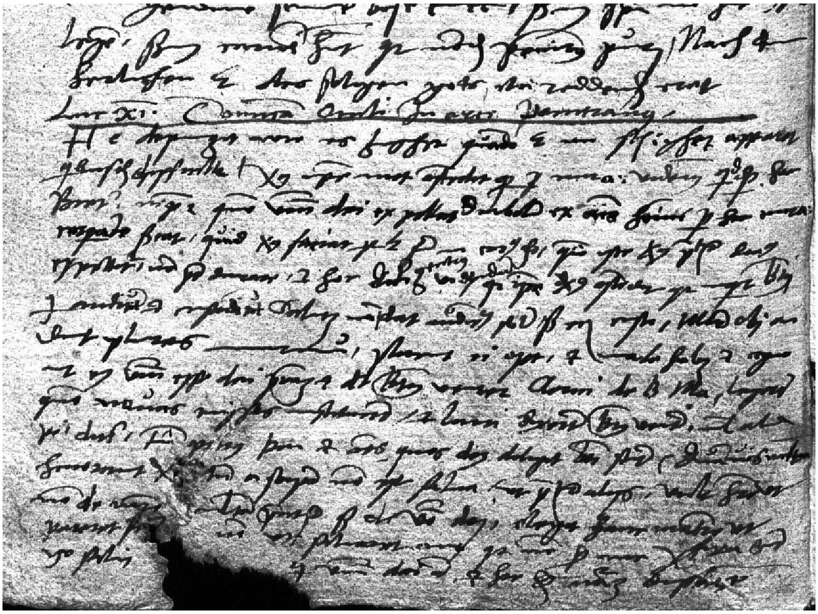


Abbildung 2. Ausschnitt aus Abb. 1. Unterhalb der Linie die abgekürzten Worte: *Hoc evangelium depingit, wie es zughet, quando evangelium im schwang ghet, apparet quibusdam difficile. Christus […].*

Briefe, Tischreden etc. Luthers aufzuzeichnen. Die Arbeitsweise Rörers lässt sich sehr gut in den Manuskripten erkennen. Der überwiegende Teil der Handschriften ist durch eine sehr kleine, flüchtige und extrem stark gekürzte Kursive gekennzeichnet, in der zwischen Latein und Deutsch unvermittelt gewechselt wird (Abb. 2). Die abgekürzten Worte sind teilweise so fragmentarisch reduziert, dass eine korrekte und auch eindeutige Auflösung oft nur schwer möglich ist. Paul Pietsch hat eigens eine Übersicht der häufigsten von Rörer verwendeten Abkürzungen vorgelegt (Pietsch 1898 IV–VII und Pietsch 1904 XVII–XXIV), von einer eigenständigen und systematischen Kurzschrift zu sprechen, wäre allerdings nicht zutreffend. Dass selbst Rörer seine Aufzeichnungen für schlecht nachvollziehbar hielt, zeigen seine eigenhändigen Überarbeitungen von Predigten, in denen er nachträglich die unleserlichsten Stellen ausbesserte. Als 1537 die Ausgabe von Lutherwerken anhand der Rörer-Handschriften begonnen werden sollte, scheiterte das Projekt vorerst, da offensichtlich niemand willens und in der Lage war, Rörers Handschrift zuverlässig zu entziffern. Auch der Nürnberger Ratsherr Hieronymus Baumgärtner musste 1547 gestehen, dass er die ihm zur Verwahrung gesandten Handschriften Rörers nicht lesen könne (Flemming 31, Anm. 2).

Kurfürst Johann Friedrich I. entband Georg Rörer 1537 von den Pflichten seines geistlichen Amtes in Wittenberg und ernannte ihn zum Adlatus Luthers, um die Drucklegung der Luther-Bibel und anderer Schriften des Reformators voranzutreiben. Durch den Tod Luthers 1546 und die politischen Ereignisse in Mitteldeutschland (Schlacht von Mühlberg 1547) gerieten die Arbeiten in Wittenberg ins Stocken. 1553 setzte man dann in Jena mit der gleichnamigen Ausgabe die Publikation lutherischer Schriften fort. Als wichtigste Grundlage sollten dabei die 35 Handschriftenbände Rörers dienen. Kurz vor seinem Tod übertrug Rörer die Besitzrechte seiner Sammlung, zu der auch drei Drucke, die sogenannten Handexemplare des Alten (VD 16 B 2704) und des Neuen Testaments (VD 16 B 4429) sowie Rörers Korrekturexemplar eines Bandes der deutschen Reihe der Wittenberger Lutherausgabe (VD 16 L 3316) gehörten, den Ernestinern. Nach Rörers Tod in Jena 1557 wurde sein »Nachlass« der hiesigen Bibliothek, die 1558 zur Universitätsbibliothek erhoben wurde, übergeben. Hier, in der heutigen ThULB, befinden sie sich noch heute (vgl. zu den Handschriftenbeständen Ott 49–54 sowie die Einleitung in Klein-Ilbeck/Ott). Mit Ausnahme einiger nicht von Rörer stammenden Autographen und einem Band, der von Michael Stifel geschrieben wurde, sind alle Manuskripte des sogenannten Nachlasses von der Hand Rörers. Der Jenaer Bestand bildet aber nicht die Gesamtheit des Rörerschen Erbes. Der Verbleib von ca. 50 Drucken, die Rörer noch zu Lebzeiten besessen haben soll, ist unklar, ebenso sind Handschriftenkonvolute verloren gegangen (Volz/Wolgast 175 f., 183, 186 und 274). Des Weiteren gelangte eine Foliohandschrift in die Staats- und Universitätsbibliothek Hamburg (Sup. ep. [2°] 92, vgl. Volz/Wolgast 263–274) und Teile einer Quarthandschrift in die Herzog August Bibliothek nach Wolfenbüttel (214 Gud. Lat. 4°, vgl. Volz/Wolgast 158). In Manuskripten der Forschungsbibliothek Gotha befinden sich Teilabschriften aus Rörer-Bänden (B 168, vgl. Volz/Wolgast 1970, S. 65 und 234, Anm. 3).

3 Projektbeschreibung

Um diesen verstreuten Bestand auf einer Online-Plattform zugänglich zu machen und die Rörer-Sammlung erstmals als einen Bestand *sui generis* zu erschließen und um außerdem die Publikationsformen des »Digitalen Zeitalters« anzuwenden, wurde von der ThULB und dem Lehrstuhl für Kirchengeschichte der Universität Jena ein gemeinsames Forschungsprojekt bei der Deutschen Forschungsgemeinschaft (DFG) beantragt und im Frühjahr 2008 auch von dieser genehmigt. Das Erschließungsprojekt ist auf vier Jahre konzipiert. In der ersten Projektphase werden alle formalen und inhaltlichen Merkmale der Jenaer Handschriften im Sinne einer klassischen Handschriftenbeschreibung, wie sie traditionell in gedruckte Kataloge mündet, erschlossen. Hierbei werden dieselben Kriterien und Regeln angewandt, wie sie von der DFG für die Katalogisierung mittelalterliche Handschriften empfohlen werden (Richtlinien Handschriftenkata-

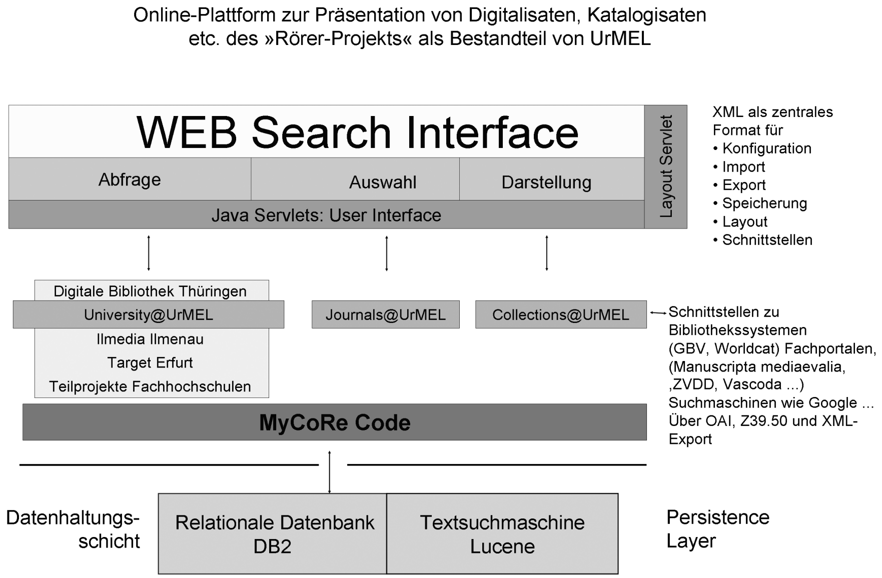

Online-Plattform zur Präsentation von Digitalisaten, Katalogisaten
etc. des »Rörer-Projekts« als Bestandteil von UrMEL
WEB Search Interface
Abfrage
Auswahl
Darstellung
Java Servlets: User Interface
Layout Servlet
XML als zentrales Format für
• Konfiguration
• Import
• Export
• Speicherung
• Layout
• Schnittstellen
Digitale Bibliothek Thüringen
University@UrMEL
Ilmedia Ilmenau
Target Erfurt
Teilprojekte Fachhochschulen
Journals@UrMEL
Collections@UrMEL
Schnittstellen zu Bibliothekssystemen (GBV, Worldcat) Fachportalen, (Manuscripta mediaevalia, ‚ZVDD, Vascoda ...) Suchmaschinen wie Google ... Über OAI, Z39.50 und XML-Export
MyCoRe Code
Datenhaltungs-schicht
Relationale Datenbank DB2
Textsuchmaschine Lucene
Persistence Layer

Abbildung 3. Struktur von UrMEL.

logisierung). Die durch die Tiefenerschließung erhobenen Daten werden direkt mittels ManuscriptumXML (HiDA4) in die Datenbank »Manuscripta Mediaevalia« eingepflegt, um sie möglichst rasch im Internet recherchierbar zu machen.

Parallel dazu wird in der ThULB die bereits bestehende Online-Plattform UrMEL (University Multimedia Electronic Library of Jena) um ein aus der Erforschung der Rörer-Sammlung letztlich erwachsendes Portal zur Luther- und Reformationsforschung erweitert (Georg-Rörer-Projekt). Dort werden ab Ende 2009 erste Katalogisate und Digitalisate zur Verfügung stehen. Im weiteren Verlauf des Projektes sollen die auf andere Bibliotheken verteilten Autographen Rörers in die Digitalisierung und Auswertung mit einbezogen und zur Verfügung gestellt werden, so dass erstmals das Arbeiten mit dem Gesamtbestand von Georg Rörers Handschriften an einem zentralen Ort möglich sein wird. Die dann somit virtuell wiedererstandene Gesamtheit der erhaltenen Manuskripte wird völlig neue Forschungsperspektiven eröffnen. Gemäß den Visionen der »eScience History« (Sahle 67) wären Quellenüberlieferung, Erschließungsmittel und Forschungsliteratur in einer fach- und themenspezifischen Arbeitsplattform integriert.

Den ausgearbeiteten Katalogisaten werden in der Online-Präsentation die vollständigen Digitalisate der Rörer-Bände zur Seite gestellt, so dass dem jeweiligen Blatt einer

Handschrift seine wissenschaftliche Beschreibung zugeordnet werden kann. Dies ermöglicht jedem interessierten Wissenschaftler, in den Katalogisaten, die wie Hypertexte strukturiert werden, zu recherchieren, sofort alle bisher nicht gedruckten Texte zu identifizieren, schwierige Lesarten zu überprüfen, nach besonderen Merkmalen zu suchen oder die bisher nicht identifizierten Texte einer eigenen Untersuchung zu unterziehen. Indices und Volltextsuche bilden dabei die Grundlagen der Recherche. Die Digitalisate (archiviert als TIFF-Dateien, 24-Bit-Farbtiefe, Auflösung mindestens 300 dpi, bereitgestellt im Format JPEG) bieten den unschlagbaren Vorteil der optischen Vergrößerung, die gerade bei Rörers Handschrift oftmals der Bildschirmarbeit den Vorzug vor der Autopsie geben.

Damit die Daten der Handschriftenkatalogisierung sowohl in »Manuscripta Mediaevalia« als auch über UrMEL recherchierbar sind, wird von Projektmitarbeitern ein Java-Programm entwickelt, das in der Lage ist, die aus ManuscriptumXML importierten Daten (XML mit eingeschobenen RTF-Blöcken) zu lesen und in reine XML-Daten umzuwandeln. Diese Daten, die auch als Metadaten bspw. von Google abfragbar sind, werden über das *content management system* MyCoRe der UrMEL-Plattform zur Verfügung gestellt und können dort über eine Suchmaske recherchiert werden. Der doppelte Nutzen dieser Maske besteht darin, dass sie als Editor zur Nachnutzung in anderen Projekten auch Daten aufnehmen kann, die bei Bedarf von ManuscriptumXML importiert werden könnten. Ein weiterer Vorteil dieser Anwendung ist, dass sowohl die importierten Daten aus ManuscriptumXML als auch die direkt über die Maske eingegebenen Werte in reine XML-Daten umgewandelt werden, was langfristig eine optimale Nutzung der Metadaten etc. sichert (vgl. die Strukturen in Abb. 3).

Um die traditionelle, aber willkürliche Trennung von Bibliotheksbeständen in Manuskripte und Druckwerke zu überwinden, werden der inneren Logik des Projektes folgend ebenso die Drucke der Rörer-Sammlung erfasst. Damit werden sie in derselben Datenbank recherchierbar sein wie die Bände der ehemaligen Wittenberger »Bibliotheca Electoralis«, einem Gründungsbestand der ThULB Jena. Die lokale wie inhaltliche Nähe dieser ehemals kurfürstlichen Bibliothek, die in einem eigenen Projekt[1] der ThULB vollständig digitalisiert und inhaltlich erschlossen wird, kann in Ergänzung zu den von Rörer selbst verfassten Handschriften und herausgegebenen Drucken den Blick auf die von Rörer nachweislich benutzten Werke und damit auf Wissensbestände des frühen 16. Jahrhunderts lenken. Erst diese als virtuelle Gesamtheit in UrMEL präsentierten Handschriften, Drucke, Katalogisate und Digitalisate werden es ermöglichen, die Rezeption von Texten und damit Wissen, deren Reproduktion, Interpretation und/oder Umwandlung als Einheit, zum Beispiel im Sinne einer »Ideengeschichte«, zu erforschen. Letztlich ist es ein Anliegen des Projekts, die Rolle und Funktion Rörers,

[1] Siehe ggf. weiterführend die Notiz zum Projekt »Wissenschaftliche Aufarbeitung, Digitalisierung und Präsentation der Bibliotheca Electoralis« auf der Projekte-Seite der ThULB (DFG-Projekte). Eine eigene Projekt-Startseite wird in Kürze aufgebaut werden.

respektive seiner Sammlung, im Kontext der Reformation neu zu bewerten und der Rezeptionsforschung zu den enthaltenen Texten sowie der Reformationsforschung neue Impulse zu verleihen.

Im Unterschied zum klassischen Handschriftenkatalog, der oft erst nach Jahren »stiller Arbeit« als unveränderbarer Druck an die Öffentlichkeit gelangt, sollen die Katalogisate des Rörer-Projekts möglichst sukzessive und ohne Zeitverzögerung bewusst als *work in progress* präsentiert werden, um die Auseinandersetzung mit diesen für die Reformationsforschung wichtigen Handschriften – gerade jetzt während der »Lutherdekade« – zu fördern. Dabei sollen Forschungsergebnisse, die durch Nutzer der Datenbank gewonnen werden, schnellstmöglich in die Katalogisate aufgenommen und die Autorschaft eines Jeden kenntlich gemacht werden.

Die freie Verfügbarkeit von Handschriften und Katalogisaten bringt aber auch Schwierigkeiten mit sich, die noch nicht abschließend gelöst sind. So birgt zum Beispiel die massenweise Online-Präsentation von Digitalisaten immer die Gefahr des *information overflow* in sich. Eine Online-Datenbank zu besuchen ist eben nicht dasselbe wie der Besuch in einer Handschriftenabteilung oder in einem Archiv. In letzteren gelangt man meist erst über mehrere Stufen der Recherche und persönlichen Kontaktaufnahme zu den eigentlichen Quellen. Diese allmähliche Annäherung führt dazu, dass Informationen zu den gesuchten Materialien, wie Überlieferungslage, Forschungssituation, Editionsstand etc. bereits vorstrukturiert werden. Auch die Frage, welche Personen oder Institutionen sich ebenfalls mit den gesuchten Manuskripten beschäftigen, kann oft erst vor Ort beantwortet werden. Vor diesem Hintergrund versucht das Jenaer Projekt den Nutzern der Online-Ressourcen mehr Orientierung und eine Moderation der Forschung bzw. des Diskurses zu bieten, indem Digitalisate nicht ohne Katalogisate präsentiert und Kontakte zwischen Wissenschaftlern vermittelt werden sollen. Letzteres könnte durch eine freiwillige Nutzeranmeldung praktiziert werden. Welche Daten dabei erbeten werden und inwieweit eine anonymisierte Anmeldung möglich sein wird, ist noch nicht endgültig entschieden. Jedoch sollten die Informationen zum Nutzer mindestens Aussagen über Hintergrund und Ziel der Beschäftigung mit den Handschriften treffen, um bei einer Anfrage durch andere Nutzer den Kontakt zu den aktuellen Bearbeitern der Materialien zu vermitteln. Dies entspräche den Gepflogenheiten der in Handschriftenabteilungen und Archiven üblichen Anmeldeprozeduren und Benutzeranträge. Ob dieses Verfahren funktioniert, liegt in der Entscheidungsfreiheit der Nutzer. Im Idealfall könnte so eine bestmögliche Vernetzung von Forschern erreicht werden.

Bibliographie

DFG-Projekte [... der Thüringer Universitäts- und Landesbibliothek Jena]. <http://www.thulb.uni-jena.de/DFG_Projekte.html>.

Flemming, Paul. »Zum Briefwechsel Georg Rörers.« *Beiträge zur bayerischen Kirchengeschichte* 19 (1913): 27–37.

[Georg-Rörer-Projekt]. <http://www3.thulb.uni-jena.de/roerer>.

Klein-Ilbeck, Bettina und Joachim Ott. *Die Handschriften der Thüringer Universitäts- und Landesbibliothek Jena, Bd. 2: Die mittelalterlichen lateinischen Handschriften der Signaturenreihen außerhalb der Electoralis-Gruppe.* Unter Mitarbeit von Gerhardt Powitz und Bernhard Tönnies. Wiesbaden: Harrassowitz, 2009 [im Druck].

Lorentzen, Tim. *Johannes Bugenhagen als Reformator der öffentlichen Fürsorge.* Spätmittelalter, Humanismus, Reformation 44. Tübingen: Mohr Siebeck, 2008.

Ott, Joachim. »Georg Rörer (1492–1557) und sein Nachlass in der Thüringer Universitäts- und Landesbibliothek Jena«, in: *Ott/Treu* 47–57.

Ott, Joachim und Martin Treu, Hrsg. *Luthers Thesenanschlag – Faktum oder Fiktion.* Schriften der Stiftung Luthergedenkstätten in Sachsen-Anhalt 9. Leipzig: Evang. Verl.-Anst., 2008.

Pietsch, Paul, Hrsg. *D. Martin Luthers Werke. Kritische Gesamtausgabe.* Bd. 20. Weimar: Böhlau, 1898.

Pietsch, Paul, Hrsg. *D. Martin Luthers Werke. Kritische Gesamtausgabe.* Bd. 29. Weimar: Böhlau, 1904.

Richtlinien Handschriftenkatalogisierung. Hrsg. von der Deutschen Forschungsgemeinschaft, Unterausschuss für Handschriftenkatalogisierung. Fünfte erweiterte Ausgabe. Bonn-Bad Godesberg: Dt. Forschungsgemeinschaft, 1992.

Sahle, Patrick. »eScience History?« *Von Nowgorod bis London. Studien zu Handel, Wirtschaft und Gesellschaft im mittelalterlichen Europa. Festschrift für Stuart Jenks zum 60. Geburtstag.* Hrsg. von Marie-Luise Heckmann und Jens Röhrkasten. Nova Mediaevalia 4. Göttingen: V&R Unipress, 2008. 63–74.

UrMEL: *University Multimedia Electronic Library of Jena.* <http://www.urmel-dl.de>.

VD 16 B 2704: *Biblia: das ist: die gantze Heilige Schrifft Deudsch. D. Mart. Luth.*, Wittenberg: Hans Lufft 1538/39 (ThULB: Ms. App. 24).

VD 16 B 4429: *Das Newe Testament. D. Mart. Luth.*, Wittenberg: Hans Lufft 1540 (ThULB: Ms. App. 25).

VD 16 L 3316: *Der Ander Teil der Bücher D. Mart: Luth.*, Wittenberg: Georg Rhau (Erben) 1551 (ThULB: Ms. App. 26).

Volz, Hans und Eike Wolgast, Hrsg. »Beschreibendes Verzeichnis der in unserer Ausgabe Briefe angeführten Lutherbriefhandschriften.« *WA. Briefwechsel.* Bd. 14. Weimar: Böhlau, 1970. 175–282.

WA: *D. Martin Luthers Werke. Kritische Gesamtausgabe.* Weimar: Böhlau, 1883 ff.

Codicological Descriptions in the Digital Age*

Timothy Stinson

Abstract

Although some of the traditional roles played by codicological descriptions in the print era have not changed when translated to digital environments, other roles have been redefined and new ones have emerged. It has become apparent that in digital form the relationship of codicological descriptions to the books they describe has undergone fundamental changes. This article offers an analysis of three of the most significant of these changes: 1) the emergence of new purposes of and uses for these descriptions, especially with respect to the usefulness of the highly specific and specialized technical language common to codicological descriptions; 2) a movement from a one-to-one relationship between a description and the codex that it represents to a one-to-many relationship between codices, descriptions, metadata, and digital images; and 3) the significance of a shift from the symmetry of using books to study other books to the asymmetry of using digital tools to represent and analyze books.

Zusammenfassung

Einige der traditionellen Funktionen kodikologischer Beschreibungen aus dem Druckzeitalter haben sich im Übergang in eine digitale Umgebung nicht verändert. Andere Funktionen aber sind neu definiert worden oder überhaupt erst entstanden. Offensichtlich hat sich im Digitalen das Verhältnis zwischen den kodikologischen Beschreibungen und den Büchern, die sie beschreiben, fundamental gewandelt. Dieser Beitrag untersucht drei der wichtigsten Veränderungen: 1) die Entstehung neuer Zwecke und Verwendungsweisen dieser Beschreibungen, insbesondere in Bezug auf die Nützlichkeit des sehr speziellen Fachvokabulars, das in kodikologischen Beschreibungen üblich ist; 2) die Entwicklung von einer 1:1-Beziehung zwischen einer Beschreibung und dem dadurch sie repräsentierten Codex zu einer 1:n-Beziehung zwischen Codizes, Beschreibungen, Metadaten und digitalen Abbildungen; und 3) die Bedeutung des Übergangs von der Symmetrie, Bücher zu benutzen, um andere Bücher zu untersuchen, zu der Asymmetrie, digitale Werkzeuge zu nutzen, um diese Bücher wiederzugeben und zu analysieren.

[*] The author wishes to thank his colleagues at the *Roman de la Rose* Digital Library for their feedback on drafts of this article. Mark Patton, Beatrice Radden Keefe, and David Reynolds lent their expertise and provided both needed criticism and encouragement during various stages of the writing and revision process.

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 35–51.

1 Introduction

It is by now well established that electronic technologies have fundamentally altered the form and uses of critical editions, as attested in electronic editions and texts produced by scholars such as Hoyt Duggan, Jerome McGann, and Kenneth Price, and in articles and anthologies—by these scholars and others—that document the scope and importance of this shift.[1] More broadly, these editions and critical discussions have made manifest the fundamental impact that technologies such as electronic databases, hypertext, digital imaging, and tools for searching and manipulating texts have had on scholarly practices and the production, use, and reception of cultural artifacts. The goal of this essay is to articulate the significance of these developments to the authoring and use of codicological descriptions, a genre that has received comparatively little attention in these discussions. While my observations are applicable to any codicological descriptions available in electronic form—and frequently to digitized bibliographical descriptions and other forms of analytical descriptions and catalog records—I will focus primarily on digitized manuscript descriptions from the *Roman de la Rose* Digital Library (RRDL) and the Parker Library on the Web project as case studies to illustrate my points.

The aim of the RRDL, a joint project of Johns Hopkins University and the Bibliothèque nationale de France, is to provide digital surrogates of all extant manuscripts of the *Roman de la Rose*, a 13th-century poem surviving in more than 300 manuscript copies. Surrogates are accompanied by and linked to full codicological descriptions, as there are many features of physical books that may be inaccessible or unclear when represented in digital form. The manuscripts are held by a wide variety of local, national, and university libraries, art museums, and private collectors worldwide, and as a consequence only brief or provisional descriptions of them exist in many cases. There is no comprehensive catalogue or other reference work in print containing descriptions or even a complete list of *Rose* manuscripts; prior to the RRDL, the most recent such work was Ernest Langlois's *Les Manuscrits du Roman de la Rose*, which was published in 1910 and offers short descriptions of approximately two-thirds of *Rose* manuscripts known to survive today. Because of this, the RRDL has undertaken the task of writing new descriptions for the site, a process in which I have been actively involved during the past several years. The Parker Library on the Web project, meanwhile, is in the process of digitizing the holdings of one collection—the famous library assembled

[1] Duggan is the Project Director of the Society for Early English and Norse Electronic Texts and co-editor of several hypertext editions of Piers Plowman manuscripts published by the *Piers Plowman Electronic Archive*, which he also directs. McGann is the editor of *The Complete Writings and Pictures of Dante Gabriel Rossetti: A Hypermedia Archive* and a leader on a number of collaborative digital projects, including the Networked Infrastructure for Nineteenth-Century Electronic Scholarship. Price is co-editor of *The Walt Whitman Archive*.

in the sixteenth century at Corpus Christi College, Cambridge, by Matthew Parker, a well-connected book collector and public figure who played a key role in the English reformation. The Parker project inherits a rich tradition of descriptions of the library's collection, which has been catalogued four times,[2] including the rather thorough work of M. R. James, whose *A Descriptive Catalogue of the Manuscripts in the Library of Corpus Christi College Cambridge* was published in two volumes—the first in 1909 and the second in 1912. The James *Catalogue* is being digitized by the Parker project, with descriptions marked up and linked to the medieval codices they describe.

In many ways, the descriptions that I have written for the RRDL differ little from traditional work of this sort; I analyze the physical books in person and produce a prose description that, in print form, would not be out of place in catalogues of manuscripts produced twenty-five or even a hundred years ago. The descriptions on the Parker site, meanwhile, are marked up versions of work begun over a century ago by James. Yet in marking up both sets of descriptions—one custom made for the web, the other a digitized version of a printed reference work—for inclusion in digital libraries, and in designing and implementing interfaces for accessing XML-encoded descriptions and the surrogates to which they are linked, it has become apparent that in digital form the relationship of codicological descriptions to the books they describe has, like the relationships of critical editions to the texts they document and represent, undergone fundamental changes. I will offer here an analysis of three of the most significant of these changes: 1) the emergence of new purposes of and uses for these descriptions, especially with respect to the usefulness of the highly specific and specialized technical language common to codicological descriptions; 2) a movement from a one-to-one relationship between a description and the codex that it represents to a one-to-many relationship between codices, descriptions, metadata, and digital images; and 3) the significance of a shift from using "books to study books"—in this context, printed codices containing descriptions that represent other codices—to hypertext descriptions that escape "the time-and-space frames established by the material characteristics of the book" (McGann 20, 22).

2 Evolution in the Purposes and Uses of Codicological Descriptions

Some of the traditional roles played by printed codicological descriptions have not changed in digital environments. Descriptions formalize an approach to and vocabulary for understanding cultural artifacts, and they provide an expert opinion on the origins and status of manuscript books for the benefit of scholars who are unable to con-

[2] The four catalogues are by Thomas James (1600), William Stanley (1722), James Nasmith (1777), and Montague Rhodes James (1909–1912). For more information, see "About the Catalogs" on the Parker Library on the Web site.

sult the original objects and/or non-expert users who lack the necessary skills to make such judgments for themselves. Because they typically summarize dates, origins, owners, and contents of books, descriptions also serve as useful preliminary resources for researchers looking for information that will suggest which volumes, collections, and repositories are most likely to reward further time and effort. Codicological descriptions are usually characterized by a highly specialized and specific vocabulary—developed and augmented over the years by curators, codicologists, art historians, and others—and terse prose entries that make highly efficient use of space in printed books. Abbreviations and formulae are common, as they facilitate conveying a considerable amount of information in a brief space. These features are seen clearly in the following excerpt from a description of M. 948, a *Rose* manuscript held by the Morgan Library & Museum and available in surrogate form through the RRDL:

> M. 948 GUILLAUME DE LORRIS AND JEAN DE MEUN. Roman de la rose. France, about 1520, written by Girard Acarce for Francis I, king of France.
>
> Vellum, 210 leaves (10 5/16 x 7 5/16 in.) (262 x 186 mm.), foliated. 2 cols., 33 lines (180 x 125 mm.). Gothic script, black and some gold ink, written by Girard Acarce. 2 full-page miniatures with architectural frames, 67 large miniatures with full-page architectural frames which also include portions of text, 38 small miniatures (half-column) with simple gold frames, 2 small decorated borders, numerous gold initials against alternating red and blue backgrounds throughout. The miniatures are by at least two distinct artists: examples by the stronger are fols. 77v, 83v, 95; the weaker, fols. 172, 180, 186. Collation: I4, II8, III7, IV8, V2, VI8–XIII8, XIV6, XV8–XXVI8, XXVII6, XXVIII9. Binding: Modern red velvet, edges gilt and gauffered, with a row of lozenges containing the letter F flanked by rows of lozenges containing fleur-de-lis.
>
> The text is complete except for two breaks: a leaf between fols. 12 and 13 (containing lines 656–768 of M. Méon, Le roman de la rose, Paris, 1814, I, 23–32, and a small miniature probably depicting caroling or dancing), and two conjoint leaves between fols. 198 and 199 (containing lines 20907–21125 of Méon, III, 282–291, and a large miniature probably depicting Pygmalion at work).[3]

Such descriptions have traditionally met (and continue to meet) the needs of two types of users. The first is the visitor to the library who wishes to use the description as a guide to a manuscript being consulted in person. Information such as the name of the scribe responsible for the manuscript, the location of—and text lost as a result of—missing leaves, and the distribution and relative merit of the work of the two artists facilitates and expedites the work of most researchers, and is particularly valuable in enabling the work of those who wish to consult the manuscript for literary, historical, or other

[3] This description is available in hard copy to visitors of the Morgan's reading room and in PDF via *CORSAIR*, the Morgan's online catalogue. It continues with a detailed list of the subject matter of the miniatures.

reasons, but are not themselves equipped with the specialized knowledge to make such judgments. The second type of user whose needs are met by such descriptions is the researcher who is studying the manuscript remotely, and thus needs information that would otherwise be available only if the manuscript were at hand. The fact that the manuscript contains two columns of 33 lines each, for example, or that there are "numerous gold initials against alternating red and blue backgrounds throughout" is information that one does not need to provide to a library visitor who has the manuscript in front of her. In such cases, the language of codicological descriptions has needed to be precise and clear because it needed to convey an image of an original object that a user often could not see in person.

In digital environments, we encounter new forms for both codicological descriptions and the objects they describe. As Daniel Pitti has observed, "[i]n order to apply computer technology to humanities research, it is necessary to represent in machine-readable form the artifacts or objects of primary or evidentiary interest in the research, as well as secondary information used in the description, analysis, and interpretation of the objects" (474). In digital libraries such as the RRDL and Parker Library on the Web, color digital images of manuscript codices are the machine-readable forms of the original artifacts, and XML-encoded codicological descriptions are the secondary information used to describe, analyze, and interpret these artifacts. Some purposes and uses of these descriptions—and the precise, specialized language used to write them—remain the same or very similar to their print predecessors, while others are being transformed as a result of digitization. In order to demonstrate this, I will focus here on three categories of information found in hypertext codicological descriptions, as well as their similarities to and departures from their print analogues. The first of these—the dissemination of specialized knowledge—remains relatively unchanged; whether working in a physical or an online library, many users will need the combined paleographical, codicological, literary, and art historical knowledge found in descriptions such as that of M. 948 above. Such information is of course more easily searched, mined, and disseminated in a digital environment, but this is true much more broadly of marked up texts of all types, and thus need not detain us here. The second category is information that refers to the physical nature of manuscripts, and hence is not available via digital surrogates. This includes physical measurements of bindings, folios, and text blocks, tactile information such as the thickness of paper or whether one is seeing the hair or flesh side of parchment, and a reliable collation of the book, which necessitates physical inspection. This category of information also serves as a check against distortions to our understanding of physical objects that occur in electronic environments. Online libraries and archives are frequently equipped with tools for manipulating images, such as the ability to zoom, pan, and rotate; for example, the RRDL allows users to choose three display sizes in order to accommodate the variety of monitors which visitors to the online library may be using, and adds to this a larger "popup" option and a number of zoom and pan tools. While such technologies are enormously useful to a researcher wanting to conduct a detailed analysis of a miniature or marginal inscription, they also tend to distort a sense of scale, both within one book and between multiple books. Digital repositories, meanwhile, are subject to mistakes that look remarkably similar to those made centuries ago in scriptoria and binderies. Instead of mistakes in foliation or pagination, files are misnamed. A break in a digital codex might as easily be the result of a lost file as a lost leaf in the physical book it represents. And rather than a binder misordering his gatherings, we might find files sequenced incorrectly. Descriptions made from physical books therefore serve as a means to diagnose and correct such problems.

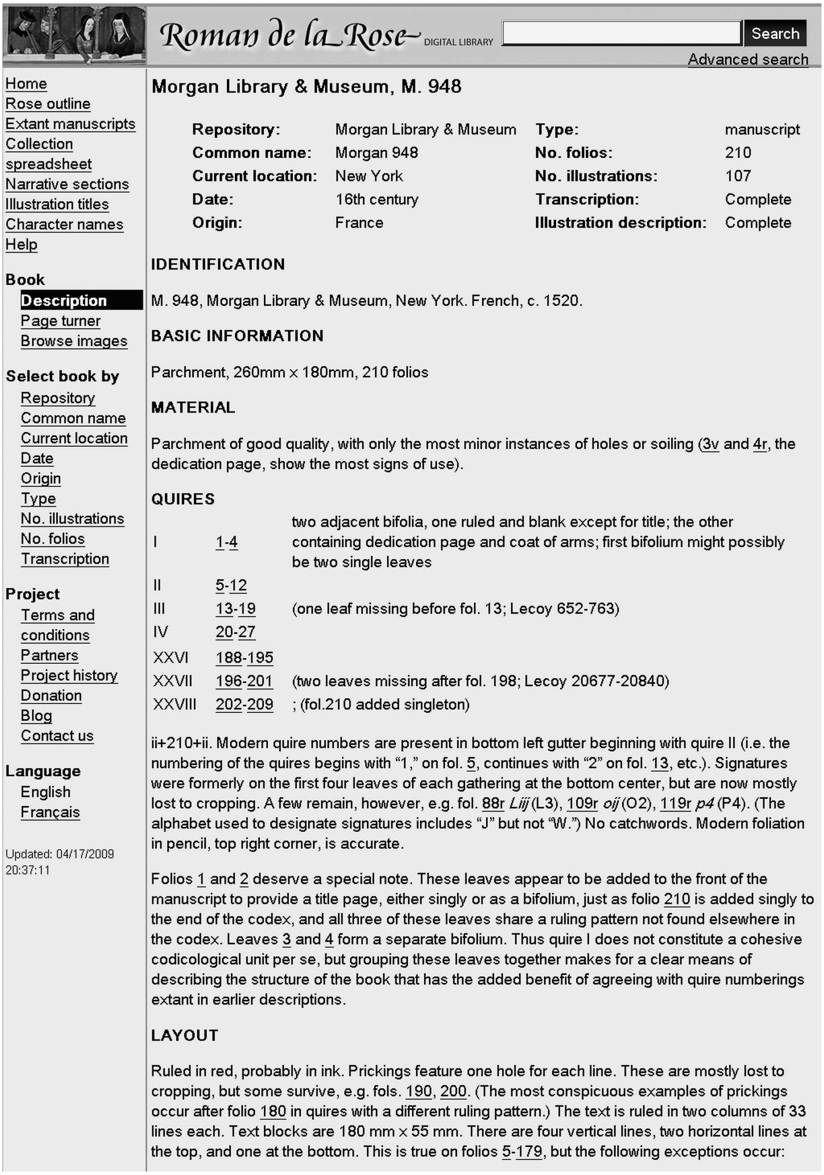
The third category pertains to information that previously was included to meet the needs of those researchers studying manuscripts remotely through descriptions of them; it is in this category that we witness the most fundamental changes in the purposes and uses of codicological descriptions in digital environments. This category concerns information that, in printed descriptions, was designed to summarize and provide details of the physical appearances of manuscripts. Needless to say, the need for such information is substantially lessened when descriptions are accompanied by digital images; that the text of M. 948 is in two columns or that there are "numerous gold initials against alternating red and blue backgrounds throughout" is now attested by the images themselves, and thus there is not the same need for this information in the description. But this information has gained new usefulness even as it has lost much of its original purpose, for it now serves as a means for sorting, classifying, and comparing collections of manuscripts:

> Most historical or traditional documents and records are too irregular for direct representation in databases. Data in databases are rigorously structured and systematic and most historical documents and records simply are not. [...] While database technology may be inappropriate for representing most historical documents and records, it is very appropriate technology for recording analytic descriptions of artifacts and in systematically describing abstract and concrete phenomena based on analysis of evidence found in artifacts. Analytic descriptive surrogates will be useful in a wide variety of projects. Archaeologists, for example, may be working with thousands of objects. Cataloguing these objects involves systematically recording a vast array of details, frequently including highly articulated classification schemes and controlled vocabularies. Database technology will almost always be the most appropriate and effective tool for collecting, classifying, comparing, and evaluating artifacts in one or many media. (Pitti 476-77)

The textual materials comprised in the RRDL project provide clear examples of both types of documents mentioned by Pitti. Full transcriptions of manuscript versions of the 13th-century poem, which typically exceed 17,000 lines in length, are clearly "too irregular for direct representation in databases." Yet the precision and specificity of the language of codicological descriptions, developed to convey a substantial amount of information in a small space (a necessity in print reference works if one wishes to avoid prohibitive cost and unwieldy volumes) now facilitates databases that provide highly flexible, searchable, and sortable relationships between the original artifacts. When, for example, a visitor to the RRDL views the codicological description of M. 948 written expressly for the digital library, he first sees a header containing information including features such as the date, origin, number of folios, and number of illustrations, as seen in Figures 1 and 2.

This information not only summarizes data about the physical book and the RRDL's work on it (for example, we see here that a transcription of the text and descriptions of the illustrations have been completed), but provides data for a database that allows users to sort other books using these categories. A menu on the left of the screen mirrors these categories and allows one, for example, to look for all volumes from the same country, repository, or century, or to produce a list showing the numbers of folios and illustrations that will help to convey how any given manuscript fits into the spectrum of available manuscripts.[4] Does it have relatively more or fewer folios than most other manuscripts? Is the number of illustrations unusually high or low? For example, a search reveals that M. 948 is near the top of the list in both categories (and the fact that it is a relatively lengthy book with numerous miniatures surely relates to its status as a luxury copy designed for presentation to François I of France). A more complete database built upon information marked up in the codicological descriptions—including the data above plus additional information such as height, width, number of leaves per gathering, and average number of lines per column—can be viewed online or downloaded in spreadsheet format so that users of the RRDL can search, sort, and analyze this information across the entire corpus of manuscript descriptions. The specialized terms found in the descriptions that are less easily adapted to spreadsheet categories and conventions of standardization, meanwhile—from *gauffered* edges to *cursiva formata* script—are rendered searchable across the collection. And while some information is rendered less useful for basic descriptive purposes in the presence of digital facsimile images, it gains new usefulness because it is searchable across multiple descriptions; a user might no longer need a description to inform her that a miniature has a gold leaf background, but the presence of this information in hypertext descriptions means that she may now search for all other manuscripts that feature such decoration and map out other similarities (or differences) that the volumes may share.

[4] It should be noted that both the heading for M. 948 and my description of the layout of the site are a snapshot of a particular moment—April 2009—in the RRDL's evolution. These features and their placement are likely to change over time as development work continues.



Morgan Library & Museum, M. 948

Repository:	Morgan Library & Museum	**Type:**	manuscript
Common name:	Morgan 948	**No. folios:**	210
Current location:	New York	**No. illustrations:**	107
Date:	16th century	**Transcription:**	Complete
Origin:	France	**Illustration description:**	Complete

IDENTIFICATION

M. 948, Morgan Library & Museum, New York. French, c. 1520.

BASIC INFORMATION

Parchment, 260mm x 180mm, 210 folios

MATERIAL

Parchment of good quality, with only the most minor instances of holes or soiling (3v and 4r, the dedication page, show the most signs of use).

QUIRES

I	1-4	two adjacent bifolia, one ruled and blank except for title; the other containing dedication page and coat of arms; first bifolium might possibly be two single leaves
II	5-12	
III	13-19	(one leaf missing before fol. 13; Lecoy 652-763)
IV	20-27	
XXVI	188-195	
XXVII	196-201	(two leaves missing after fol. 198; Lecoy 20677-20840)
XXVIII	202-209	; (fol.210 added singleton)

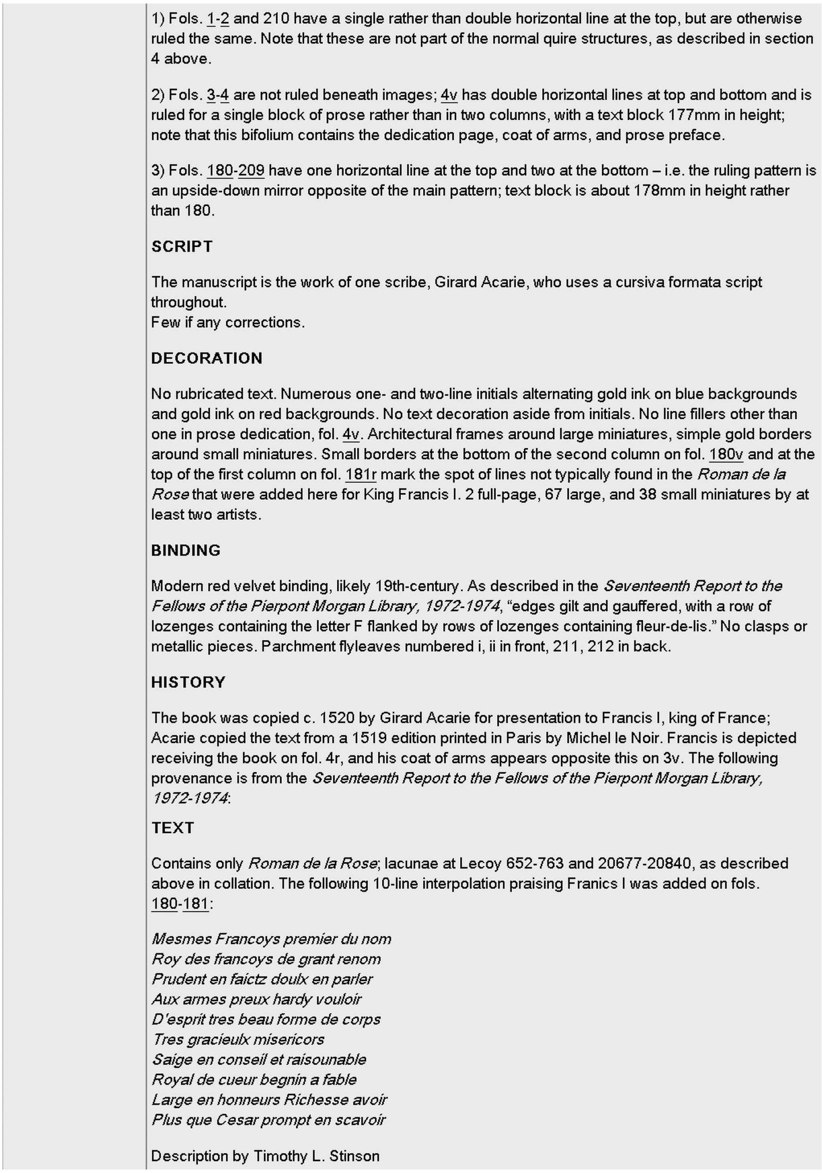
ii+210+ii. Modern quire numbers are present in bottom left gutter beginning with quire II (i.e. the numbering of the quires begins with “1,” on fol. 5, continues with “2” on fol. 13, etc.). Signatures were formerly on the first four leaves of each gathering at the bottom center, but are now mostly lost to cropping. A few remain, however, e.g. fol. 88r *Liij* (L3), 109r *oij* (O2), 119r *p4* (P4). (The alphabet used to designate signatures includes “J” but not “W.”) No catchwords. Modern foliation in pencil, top right corner, is accurate.

Folios 1 and 2 deserve a special note. These leaves appear to be added to the front of the manuscript to provide a title page, either singly or as a bifolium, just as folio 210 is added singly to the end of the codex, and all three of these leaves share a ruling pattern not found elsewhere in the codex. Leaves 3 and 4 form a separate bifolium. Thus quire I does not constitute a cohesive codicological unit per se, but grouping these leaves together makes for a clear means of describing the structure of the book that has the added benefit of agreeing with quire numberings extant in earlier descriptions.

LAYOUT

Ruled in red, probably in ink. Prickings feature one hole for each line. These are mostly lost to cropping, but some survive, e.g. fols. 190, 200. (The most conspicuous examples of prickings occur after folio 180 in quires with a different ruling pattern.) The text is ruled in two columns of 33 lines each. Text blocks are 180 mm x 55 mm. There are four vertical lines, two horizontal lines at the top, and one at the bottom. This is true on folios 5-179, but the following exceptions occur:

Figure 1. A screenshot of the description of M. 948. Courtesy of the *Roman de la Rose* Digital Library. Modified for printing; text on quires shortened.



1) Fols. 1-2 and 210 have a single rather than double horizontal line at the top, but are otherwise ruled the same. Note that these are not part of the normal quire structures, as described in section 4 above.

2) Fols. 3-4 are not ruled beneath images; 4v has double horizontal lines at top and bottom and is ruled for a single block of prose rather than in two columns, with a text block 177mm in height; note that this bifolium contains the dedication page, coat of arms, and prose preface.

3) Fols. 180-209 have one horizontal line at the top and two at the bottom – i.e. the ruling pattern is an upside-down mirror opposite of the main pattern; text block is about 178mm in height rather than 180.

SCRIPT

The manuscript is the work of one scribe, Girard Acarie, who uses a cursiva formata script throughout.
Few if any corrections.

DECORATION

No rubricated text. Numerous one- and two-line initials alternating gold ink on blue backgrounds and gold ink on red backgrounds. No text decoration aside from initials. No line fillers other than one in prose dedication, fol. 4v. Architectural frames around large miniatures, simple gold borders around small miniatures. Small borders at the bottom of the second column on fol. 180v and at the top of the first column on fol. 181r mark the spot of lines not typically found in the *Roman de la Rose* that were added here for King Francis I. 2 full-page, 67 large, and 38 small miniatures by at least two artists.

BINDING

Modern red velvet binding, likely 19th-century. As described in the *Seventeenth Report to the Fellows of the Pierpont Morgan Library, 1972-1974*, “edges gilt and gauffered, with a row of lozenges containing the letter F flanked by rows of lozenges containing fleur-de-lis.” No clasps or metallic pieces. Parchment flyleaves numbered i, ii in front, 211, 212 in back.

HISTORY

The book was copied c. 1520 by Girard Acarie for presentation to Francis I, king of France; Acarie copied the text from a 1519 edition printed in Paris by Michel le Noir. Francis is depicted receiving the book on fol. 4r, and his coat of arms appears opposite this on 3v. The following provenance is from the *Seventeenth Report to the Fellows of the Pierpont Morgan Library, 1972-1974*:

TEXT

Contains only *Roman de la Rose*; lacunae at Lecoy 652-763 and 20677-20840, as described above in collation. The following 10-line interpolation praising Franics I was added on fols. 180-181:

Mesmes Francoys premier du nom
Roy des francoys de grant renom
Prudent en faictz doulx en parler
Aux armes preux hardy vouloir
D'esprit tres beau forme de corps
Tres gracieulx misericors
Saige en conseil et raisounable
Royal de cueur begnin a fable
Large en honneurs Richesse avoir
Plus que Cesar prompt en scavoir

Description by Timothy L. Stinson

Figure 2. Screenshot continued; text on history shortened.

3 The Relationship of Description to Codex

In addition to changes in the purposes and uses of codicological descriptions, their relationship to what they describe has changed in number and complexity. Printed codicological descriptions exhibit a one-to-one relationship to the manuscript books they describe, offering a summary and analysis of the book's physical and textual properties. This is not meant to imply, of course, that the descriptions themselves or their relationships to the codices they describe are in every case simple. On the one hand, many descriptions comprise little more than a relatively brief summary of facts about a book's physical makeup and history. When many catalogues of manuscript descriptions were created, their primary goal was simply to compile a basic record of a library's holdings; in many cases neither libraries nor their visitors had any reliable means to know with any reasonable degree of comprehensiveness what manuscript materials a given library held. A good example of this is the aptly titled *A Summary Catalogue of Western Manuscripts in the Bodleian Library at Oxford.* As that great library neared the dawn of the twentieth century, curators and researchers were continually beset with difficulties of knowing what manuscript books the library possessed and, upon knowing of a book, sometimes of locating it.[5] As a result, the *Summary Catalogue* was initiated with the goal of creating a master list comprising short descriptions of all of the library's western manuscripts. Thus the complete entry for MS Douce 195, a well known *Rose* manuscript held by the Bodleian, is as follows:

> **21769.** In French, on parchment: written in the second half of the 15th cent. in France: 14 X 9 ¾ in., ii + 158 leaves, in double columns: illuminated: binding, maroon leather with gold ornament, doublé (French 18th cent.).[6]
>
> '... Le rommant de la Rose' by Guillaume de Lorris and Jean de Meung: after the usual ending come, without any break, 24 lines, beginning 'Et lors quant ie fu esueillie.' There are many fine miniatures illustrating the poem, chiefly small, but larger ones are at foll. 1, 86v, 105v, 108, 152v. On fol. 1 are the joint arms of Orleans and Savoy dimidiating each other per pale.
> Now MS. Douce 195 (Vol. IV, 550)

[5] For a detailed account of the troubles caused by this situation at the Bodleian, see Andrew Clark's *The Cataloguing of MSS. in the Bodleian Library: A Letter Addressed to Members of Congregation.* Clark outlines fundamental goals, noting that "[t]he Summary Catalogue would furnish, within a few years, a complete guide to the Western MSS. of the Library" (which would include "both MSS. quite uncatalogued and MSS. imperfectly catalogued"), as well as more colorful advantages, such as that "[t]he Summary Catalogue would effect an immediate and perpetual saving of time of the staff, and avoid much heart-burning among readers" and "enable the Library to do justice between trifling and valuable MSS." (52-54).

[6] The identification of this as an 18th-century French binding is doubtful, as it is signed by C. Lewis, an English binder active primarily during the first half of the 19th century.

The relationship of this description to the original codex is direct and clear; it reports basic physical features of the book along with the text it contains, and offers brief comments on its decoration, including a coat of arms that might suggest provenance. Together, many such descriptions document the extent and individual components of the Bodleian's collection and serve as a valuable reference work for librarians and researchers. Of course other catalogues, such as James's catalogue of the Parker Library's manuscripts, feature descriptions that might span many pages, and these descriptions are often themselves highly accomplished works of scholarship. In addition to documenting basic facts about a book, they might also record textual variations, what exemplars were likely used, the dialect and identity of scribes, the identity or school of an illuminator, and the possible users and owners of a volume over time. As such, these descriptions might be small essays that account for the shifting milieux of a codex over the many centuries of its existence. A medieval codex is rarely a simple artifact, and may more accurately be thought of as an archeological site contained within a binding; as such, accurate descriptions of these objects are often very complex documents. Even such a complex description, however, stands in a one-to-one relationship to the book it describes; the description may discuss many intersections of textual transmission and/or artistic production, but it does so because the book itself manifests its own participation in those intersections.

In digital archives comprising images, however, this one-to-one relationship is supplanted by a one-to-many relationship. At a minimum, the original book, the codicological description, and the images that constitute the surrogate book each present relationships to the other two. In such an environment, the description describes not only the original book, but also the surrogate. While the original codex maintains an ultimate authority in that it possesses the ability to show whether a codicological description and/or the surrogate codex is somehow faulty or incomplete, the reality, and indeed the very goal, of most digital libraries is that far more people will use the digitized description as a guide to the surrogate book than would ever be able to use it as a guide to the original artifact. As such, a description in a digital environment should work equally well as a guide to both. The original codex and its surrogate images also participate in one-to-many relationships. The images, like the description, are a representation of the artifact; in turn, the images are described by and linked to the description. The original codex, meanwhile, stands in a set of new relationships to virtual versions of itself, the ramifications of which will be discussed further in section 3 below. But of course, this model is frequently complicated still further, as when transcriptions and other metadata offer new sets of relationships both to the original book and to other digital representations of it. In the RRDL, for example, the images are frequently linked not only to the codicological descriptions, but also to transcriptions and to descriptions of the illustrations written by an art historian. The codicological descriptions, meanwhile, serve not only as guides to original codices and surrogate images, but, because they are marked up in XML tags that define categories of data[7], they function as the foundation of databases and, in combination, as a large searchable "meta-manuscript" that contains combined data from numerous physical codices and thousands of digital images. The descriptions not only stand in complex multiple relationships to original artifacts, images, transcriptions, and other documents, then—they also stand in multiple relationships to one another. In a sense, this has always been true of a collection of manuscript descriptions. In Langlois's catalogue of *Rose* manuscripts, the collected descriptions stand in relation to one another—as well as to the books they describe—in that together they attest to the breadth and depth of manuscript traditions of one literary text; in James's catalogue, the descriptions together attest to the breadth and depth of one collection. But in that sense, every printed book is linked to all others contained in its bibliography and footnotes. In order to release and utilize these connections in printed books, however, one must create them anew each time, flipping through the pages to make connections or discern patterns. In digital form, conversely, the connections are always available, awaiting searching, sorting, parsing, and reorganizing, even in ways that—unlike the tables or indices in Langlois and James—the descriptions' authors did not intend or imagine. In short, with printed volumes of descriptions one is limited by the form of the book itself, and it is this set of limitations that forms the subject of my final section.

4 "The Rationale of Hypertext" and Codicological Descriptions

By now, anyone familiar with the scholarship of Jerome McGann, and particularly with his famous essay "The Rationale of Hypertext", will have noted the indebtedness of my argument to his. In particular, my discussion of the "one-to-one" relationship of description to original and my use of the term "meta-manuscript" are intentional echoes of McGann's argument that "the facsimile edition stands in a one-to-one relation to its original" and his depiction of the electronic *Oxford English Dictionary* (OED) as a "meta-book", respectively (20-21). I would like to turn now to an even more direct engagement with "The Rationale of Hypertext" through an analysis of how the relationship of the codicological description to the artifact it describes has changed in that formerly both tended to be in codex form, and thus to utilize similar technologies—e.g. indexes, glossaries, and concordances—whereas in a digital environment a description lacks such symmetry of form with the object it describes. In his essay, McGann articulates the many difficulties frequently encountered in using printed critical editions to study other printed books:

[7] Manuscript descriptions are marked up using standards described in the TEI Consortium's *TEI P5: Guidelines for Electronic Text Encoding and Interchange* (see module 10, "Manuscript Description").

> Brilliantly conceived, these works are nonetheless infamously difficult to read and use. Their problems arise because they deploy a book form to study another book form. This symmetry between the tool and its subject forces the scholar to invent analytic mechanisms that must be displayed and engaged at the primary reading level—e.g. apparatus structures, descriptive bibliographies, calculi of variants, shorthand reference forms, and so forth. [...] The crucial problem here is simple: the logical structures of the "critical edition" function at the same level as the material being analyzed. As a result, the full power of the logical structures is checked and constrained by being compelled to operate in a bookish format. (21)

Printed codicological descriptions are subject to the same limitations—and the same "crucial problem"—as critical editions, and those who have labored to become familiar with the "abbreviated and coded forms" (as McGann terms similar features in printed critical editions) and collational formulae of such descriptions will attest to the difficulty of their use. For an example of this, let us turn to James's work, which is in many ways a particularly good catalogue of manuscript descriptions. The following is a representative excerpt from entry 79, described as "Pontificale (London), Codex membranaceus in folio, picturis elegantissimis et omnibus literis initialibus deauratis ornatus":

> Vellum, 157/10 X 10, ff. 24 + cclix, double columns of 30 lines. Cent. xiv– in a fine upright black hand. Music on four-line stave.
>
> *Collation:* 14 (wants 1) 210 (1 canc.) 38 || 44 58–78 (+ slip after 1) 88 98 (5 is half a leaf) 108–138 (+ slip after 3) 148–198 206 218–298 (+ slip after 1) 308–348 (6–8 removed and replaced by) 35 (six) 368 (+ slip after 7) 38 (five). (James 160)

In many ways, the James *Catalogue* is an exception that proves the rule. It would be a simple (if not quite fair) enough matter to quote many catalogues comprising terse summaries containing little more than what I excerpt here from James, and to use those examples to point towards the limitations of printed collections of manuscript descriptions. But this would point only to one limitation, namely the expense and unwieldiness that result from taking up extra space in printed reference volumes, pressures that have limited the scope and shaped the language of almost all reference works made available in the form of printed books. James's descriptions, however, are particularly detailed and generous, commonly running several pages per manuscript codex and frequently containing lengthy lists of texts and illustrations accompanied by observations on aspects such as the quality of the artwork, previous scholarly uses or mentions of the manuscript, and summaries of subjects covered in miscellanies. Even so, the James *Catalogue* contains and relies upon "abbreviated and coded forms" such as those above that demand considerable expertise from the catalogue's users and owe their form—at least in part—to space-saving abbreviations and notational devices developed by James's predecessors.

More to the point, however, is that James's admirable undertaking cannot escape the limitations forced upon it by its own bookishness. These are precisely the limitations of printed critical editions articulated by McGann: James's *Catalogue* is a book form designed to study other book forms that demonstrates symmetry between the tool and subject manifested in "analytic mechanisms that must be displayed and engaged at the primary reading level" (McGann 21). In its printed form, the true power of James's work lies latent; the data is there, but it is contained in a medium that limits its utility. In order to unleash this potential, the work needs digitization, a means of eclipsing the constraints of its codex format:

> Computerization allows us to read 'hardcopy' documents in a nonreal, or as we now say a 'virtual', space-time environment. This consequence follows whether the hardcopy is being marked up for electronic search and analysis, or whether it is being organized hypertextually. When a book is translated into electronic form, the book's (heretofore distributed) semantic and visual features can be made simultaneously present to each other. A book thus translated need not be read within the time-and-space frames established by the material characteristics of the book. If the hardcopy to be translated comprises a large set of books and documents, the power of the translational work appears even more dramatically, since all those separate books and documents can also be made simultaneously present to each other, as well as all the parts of the documents. (McGann 22)

Thankfully, we do not have to hypothesize about the virtues of digitizing the James *Catalogue*, for that work is well under way as part of the work of the Parker Library on the Web project. As digital images of the Parker Library's manuscripts are made available online, they are accompanied by and linked to the text of the James *Catalogue*, which is available both in marked up form on the site and via PDF files that visitors can download. The result is that the digitized entries of both volumes of the *Catalogue* are "simultaneously present to each other", freeing James's work from the constraint of the codex and enormously facilitating its usefulness as a tool for researching the cultural heritage of the Parker Library, whether one is working with physical codices in Cambridge or virtual books on the web. In its original form, the James *Catalogue* utilized technologies—including lists, tables, and indices—not at all dissimilar to those found in the medieval books it described. In digitized form, this symmetry is eclipsed, and the result is a far more flexible and powerful tool.

As a footnote to this discussion, it is worth noting that James's collational formula quoted above points to the capacity of the printed book to shape—and perhaps misconstrue—our conception of the manuscript book, not merely in deploying "a book form to study another book form", but in invisibly shaping what our notion of a book is in the first place. Perhaps the chief virtue of such formulae in manuscript catalogues is brevity; they are concise yet convey the entire structure of the codex. But to a descriptive bibliographer working on printed books, brevity is not the chief aim of a collational formula, as made clear in Fredson Bower's landmark work *Principles of Bibliographical Description*:

> The collational formula and the basic description of an edition should be that of an ideally perfect copy of the original issue. A description is constructed for an ideally perfect copy, not for any individual copy, because an important purpose of the description is to set up a standard of reference whereby imperfections may be detected and properly analyzed when a copy of a book is checked against the bibliographical description. In a very rare book the evidence may not be sufficient to construct a perfect description, but it is better to aim at this perfect description, even though its collational formula may be incomplete and full of queries, than to misrepresent a book by describing only an imperfect individual copy. (113)

But of course in the world of manuscript books, there is only the "very rare book", the "imperfect individual copy". No "standard of reference" is possible in a set of one, nor can we speak of "an edition" of a manuscript book. This should serve as a caution, then, against applying the principles and practices of describing printed books too liberally to those of describing manuscript books. Browsing the range of meanings for the word *formula* in the OED (whether the printed volume or McGann's meta-book), one encounters the terms *prescription*, *rule*, and *principle*, all of which imply an ideal against which individual instances must conform or else be deemed incomplete and imperfect. But with a manuscript codex there is no abstract ideal against which to measure copies or other instances; there is only the presumed original form, which itself is a slippery notion given the number of additions, subtractions, and rebindings undergone by many manuscripts over the centuries since their inception. The only ideal the codicologist can envision is what a single manuscript book once was before, e.g., leaves were lost or physical evidence was destroyed by a binder; we may collate one text of *Roman de la Rose* against others, but we cannot collate one book against others in the ways that Bowers suggests we should collate printed books.

The rubrication, historiated initials, and foliated borders of incunables remind us that in the early days of print the concept of what a book should be was dominated by the manuscript codex. During recent centuries, the opposite is true; descriptions of manuscript books bear witness to the dominance of printing in forming our collective notion of what a book should be, and thus we have, for example, assigned them titles and expressed their structures in collational formulae that better reflect the realities of printed rather than manuscript books. As we seek to liberate our codicological descriptions from the constraints of "being compelled to operate in a bookish format," we should also bear in mind the opportunity to correct the assumption that such books operate—and should be described—in parallel with printed books. Both our tools and our mindsets need to be liberated from print if we are to achieve accurate representations of artifacts that were produced before the advent of printing. Our ideal for original artifacts—the manuscript codices themselves—is that they remain as stable and fixed in time as possible, the goals of our best curation and conservation efforts. But we should be eager to escape the fixity of our tools for working with and describing manuscript books—tools that are often byproducts of the technologies of the printed codex—and embrace instead new purposes and uses for our codicological descriptions, complex new sets of relationships between books, their surrogates, and the technologies we develop to study both, and our opportunity to move beyond the book in order to understand it better.

Bibliography

Bowers, Fredson. *Principles of Bibliographical Description.* New York: Russell & Russell, 1962.

Clark, Andrew. *The Cataloguing of MSS. in the Bodleian Library: A Letter Addressed to Members of Congregation.* Oxford: Horace Hart, 1890.

The Complete Writings and Pictures of Dante Gabriel Rossetti: A Hypermedia Archive. <http://www.rossettiarchive.org>.

CORSAIR: The Online Research Resource of The Pierpont Morgan Library. <http://corsair.morganlibrary.org>.

James, Montague R. *A Descriptive Catalogue of The Manuscripts in the Library of Corpus Christi College Cambridge.* 2 vols. Cambridge: Cambridge UP, 1912.

James, Thomas. *Ecloga Oxonio–Cantabrigiensis, tributa in libros duos, quorum prior continet catalogum confusum librorum manuscriptorum in illustrissimis bibliothecis, duarum florentissimarum Acdemiarum, Oxoniae et Catabrigiae.* London: George Bishop and John Norton (vol. 1); Arnold Hatfield (vol. 2), 1600.

Langlois, Ernest. *Les Manuscrits du Roman de la Rose: Description et Classement.* Lille: Tallandier, 1910.

McGann, Jerome. "The Rationale of Hypertext." *Electronic Text: Investigations in Method and Theory.* Ed. Kathryn Sutherland. Oxford: Clarendon Press, 1997. 19-46.

Nasmith, James. *Catalogus Librorum Manuscriptorum quos Collegio Corporis Christi et B. Mariae Virginis in Academia Cantabrigiensis Legauit Reverendissimus in Christo Pater Matthaeus Parker, Archiepiscopus Cantuariensis.* Cambridge: Woodyer, Merrill et al., 1777.

Parker Library on the Web. <http://parkerweb.stanford.edu>.

Piers Plowman Electronic Archive. <http://www.iath.virginia.edu/seenet/piers/>.

Pitti, Daniel V. "Designing Sustainable Projects and Publications." *A Companion to Digital Humanities.* Eds. Susan Schreibman, Ray Siemens, and John Unsworth. Malden, MA: Blackwell, 2004. 471-487.

Stanley, W. *Catalogus Librorum Manuscriptorum in Bibliotheca Collegiis Corporis Christi in Cantabrigia: Quos legauit Matthaeus Parkerus Archiepiscopus Cantuariensis.* London, 1722.

A Summary Catalogue of Western Manuscripts in the Bodleian Library at Oxford. 7 vols. Oxford: Clarendon Press, 1895-1953.

TEI Consortium, eds. "10. Manuscript Description." TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 1.3.0. Last updated February 1st 2009. TEI Consortium. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html>.

The Walt Whitman Archive. <http://www.whitmanarchive.org>.

Die Digitalisierung der deutschsprachigen Handschriften der Bibliotheca Palatina in der Universitätsbibliothek Heidelberg

Pamela Kalning, Karin Zimmermann

Zusammenfassung

Die Universitätsbibliothek Heidelberg stellt seit kurzem den gesamten Bestand der deutschsprachigen Handschriften der »Bibliotheca Palatina« online zur Verfügung. Der folgende Beitrag beschreibt zunächst die technischen Hintergründe dieses Digitalisierungsprojektes. Sodann zeigt er auf, welche Benutzungsfunktionen im Zusammenhang mit den Digitalisaten angelegt wurden. Anhand erster Zugriffszahlen wird schließlich erläutert, welche Chancen die Digitalisierung für die Nutzung der Handschriften in Forschung und Lehre bietet.

Abstract

The University Library of Heidelberg has digitized the complete collection of German manuscripts from the “Bibliotheca Palatina” and granted open access to these images. This contribution describes the technical background of the digitization project. It shows how the digital library can be used, in which ways it is actually used and what new opportunities digitization offers for research and teaching.

1 Projektgeschichte

In den vergangenen drei Jahren fand in der Universitätsbibliothek Heidelberg finanziert durch die Manfred-Lautenschläger-Stiftung ein Projekt statt, das die Bestände einer der wertvollsten Sammlungen deutschsprachiger Handschriften des Mittelalters und der Frühen Neuzeit vollständig online verfügbar gemacht hat. Es handelt sich um die 848 Codices Palatini germanici, den heute wieder in Heidelberg aufbewahrten Teil der »Bibliotheca Palatina« (vgl. zur Geschichte der Bibliotheca Palatina Wilken 1-272; Jammers; Mittler). Ihre Ursprünge reichen bis ins Jahr 1386, das Jahr der Gründung der Universität Heidelberg zurück (zu den Anfängen der Universität Heidelberg siehe Ritter; Moraw; Doerr u.a.; Miethke; Fuchs). Sie umfasste zum einen die universitären Bibliotheksbestände der ursprünglich eigenständigen Institutionen der Artisten und der drei

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 53–65.

höheren Fakultäten Theologie, Jura und Medizin, zum anderen die Bücher der Stiftsbibliothek in der Heiliggeistkirche und zum dritten die private Sammlung der Kurfürsten auf dem Heidelberger Schloss. Infolge der Eroberung Heidelbergs während des Dreißigjährigen Krieges durch den Feldherrn Tilly wurde die damals berühmteste Bibliothek nördlich der Alpen 1622/1623 als Kriegsbeute nach Rom abtransportiert und dort in der Vatikanischen Bibliothek aufgestellt. Durch Vereinbarungen, die während des Wiener Kongresses getroffen worden waren, kehrten die deutschsprachigen Handschriften im Jahre 1816 in ihre alte Bibliotheksheimat zurück. Bis auf 29 griechische und 16 lateinische Codices liegen alle übrigen, nicht deutschsprachigen Handschriften und sämtliche Drucke noch heute in den Tresoren der Bibliotheca Apostolica Vaticana in Rom. Demgegenüber ist der in Heidelberg verwahrte Teil dieser wertvollen und kulturhistorisch bedeutsamen Bestände durch die Digitalisierung und Bereitstellung im Netz nun einer breiten Öffentlichkeit leicht zugänglich (Bibliotheca Palatina Digital).

Im Folgenden sollen zunächst die Prinzipien und technischen Hintergründe der Heidelberger Handschriftendigitalisierung erläutert werden. Wie wurden die Objekte reproduziert, wie für die digitale Präsentation aufgearbeitet, welche Nutzungsfunktionen sind in der Präsentation angelegt? Wie wird in diesem Zusammenhang die wissenschaftliche Erschließung der Handschrift integriert? In einem zweiten Schritt soll dann aufgezeigt werden, welche Nutzungsmöglichkeiten sich auf der Basis der Digitalisate ergeben, und inwiefern diese bereits jetzt in Anspruch genommen werden. Hierzu wurde eine (nicht repräsentative) statistische Abfrage der Nutzungsdaten für die Monate November 2008 bis Februar 2009 vorgenommen, die als erste empirische Basis für eine kurze qualitative Darstellung der Chancen dient, die die Digitalisierung für Wissenschaft und Lehre eröffnet.

2 Digitalisierung und Bereitstellung der Handschriften[1]

Zur Reproduktion der Handschriften wurde ein speziell zur Digitalisierung von Handschriften entwickelter Kameratisch, der »Grazer Buchtisch« verwendet (vgl. Abb. 1). Er ermöglicht eine weitgehend kontaktlose Direktdigitalisierung der Objekte: Das Buch wird mit Hilfe eines Laserstrahls exakt positioniert, das aufgeschlagene Blatt jeweils durch den Sog einer Unterdruckeinrichtung fixiert. Dabei ist das Objektiv der Kamera durch eine spezielle Konstruktion im rechten Winkel auf das Blatt ausgerichtet, sodass Verzerrungen minimiert werden können. Das Buch muss bei diesem Vorgang nicht vollständig aufgeschlagen werden, sondern es genügt ein Öffnungswinkel von wenig mehr als 90 Grad, da die über dem Buch schwebende Kamera beweglich ist. Die Buchseiten werden einzeln abfotografiert, und zwar jeweils zunächst alle Recto- und an-

[1] Die folgende Darstellung zu den technischen Fragen und Benutzungsfunktionen orientiert sich stark an Effinger, Krenn und Wolf.

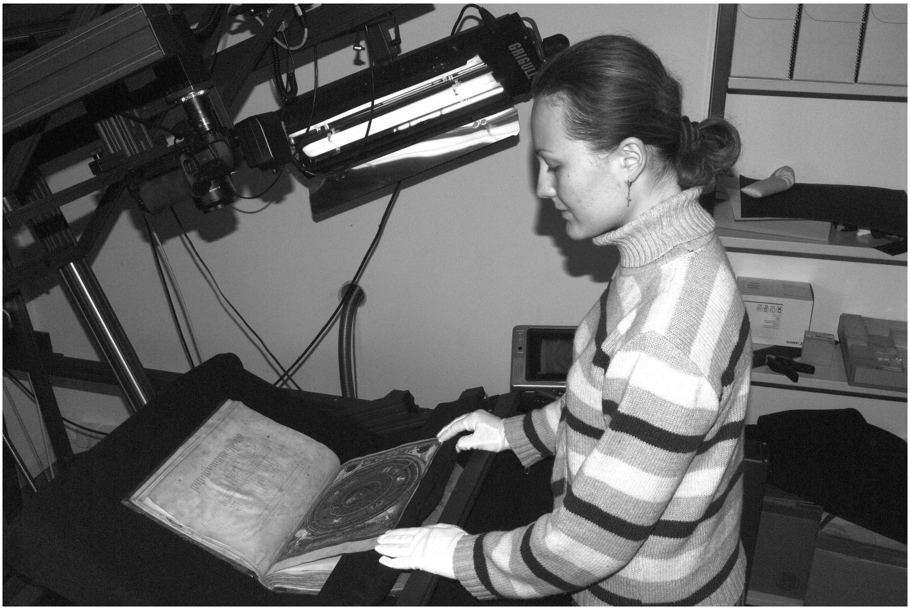


Abbildung 1. »Grazer Buchtisch« in der Digitalisierungswerkstatt der UB Heidelberg.

schließend alle Verso-Seiten, sodass das Buch während des gesamten Prozesses nur einmal gedreht werden muss. Die digitalen Bilder werden per Firewire-Schnittstelle an einen angeschlossenen PC übertragen und ohne lokale Zwischenspeicherung auf dem Festplattensystem eines Fileservers abgelegt. Dies geschieht im kameraspezifischen Rohdatenformat, um Detailverluste, Farbverfälschungen o.ä. zu vermeiden und gleichzeitig die höchstmögliche Übertragungsgeschwindigkeit zu erzielen. Für die Organisation der mit der Digitalisierung, Präsentation und Archivierung zusammenhängenden Arbeitsschritte wurde von der IT-Abteilung der UB Heidelberg ein eigenes Programm (»DWork – Heidelberger Digitalisierungsworkflow«) entwickelt. Über eine Web-Applikation wird mit dessen Hilfe einerseits die Generierung der Präsentationen, andererseits das Langzeitarchivierungssystem der Scans und der Metadaten gesteuert. Das Programm unterstützt und automatisiert sämtliche Einzelschritte von der Metadatenerstellung über die Reproduktion und Datenspeicherung bis hin zur Erstellung der Webpräsentation des einzelnen Buches. Berücksichtigung fanden bei der Programmerstellung die »Praxisregeln im Förderprogramm Kulturelle Überlieferung« der DFG sowie die »Empfehlungen der DBV AG Handschriften / Alte Drucke zur Herstellung, Internetpräsentation und Verwaltung von Digitalisaten alter Drucke und Handschriften.«

Nach der Digitalisierung werden die Bilder in das dem technischen Standard für die Langzeitarchivierung entsprechende TIFF-Format umgewandelt und mittels der Bildbearbeitungssoftware Adobe Photoshop so nachbearbeitet, dass Farb-, Helligkeits-, Kontrast- und Schärfegrad so weit wie möglich dem Original entsprechen. Die getrennt aufgenommenen Recto- und Verso-Seiten werden maschinell umbenannt und ineinander sortiert. Zur Kontrolle von Vollständigkeit und Qualität der Digitalisate werden die Bilder der gesamten Handschrift am Bildschirm durchgeblättert und überprüft; fehlende oder den Qualitätsansprüchen nicht genügende Seiten werden unmittelbar nachdigitalisiert und eingefügt. Aus den digitalen Seiten wird unter Verwendung des »Metadata Encoding and Transmission Standard« (METS) sodann das Präsentationsmodell eines virtuellen Buches erstellt. Anschließend erfolgt die Eingabe der sich am wissenschaftlichen Katalogisat orientierenden Strukturdaten, mit deren Hilfe der Text gegliedert wird. Im nächsten Schritt werden die beim Scannen erzeugten Einzeldateien in das Workflow-Programm »DWork« eingelesen und die Seiten einzeln benannt, wobei sowohl die Seiten- als auch Blattbezeichnungen möglich sind. Die Dateibenennung kann an dieser Stelle jedoch auch nach Wunsch geändert werden, z.B. wenn es sich um im Original nicht gezählte Blätter oder solche mit Sonderzählung handelt. Es folgen die Image-Konvertierungen (Umwandlung von TIFF in JPG) und die OCR-Verarbeitung. Über eine Exportfunktion werden abschließend die für die Präsentation errechneten Images und die Metadaten im METS-Format exportiert und an das auf dem Webserver der UB liegende Präsentationssystem übergeben sowie das Kopieren der Dateien zur Archivierung angestoßen.

Auf der Datenebene existiert am Ende dieses Prozesses eine für die langfristige elektronische Archivierung geeignete XML-Datei, die auch die bibliographischen Metadaten enthält. Neben diesen reinen Erschließungs-Metadaten, die im »Metadata Object Description Schema« (MODS) eingebettet werden, enthält die XML-Datei auch die Strukturdaten für die Navigation in der Handschrift. Zum Datenaustausch per OAI-Schnittstelle stehen die ebenfalls in das METS-Schema eingebetteten Dublin Core Beschreibungsdaten zur Verfügung. Da die Dateien selbst nur als reines ASCII-Format gespeichert sind, haben sie einen sehr geringen Speicherbedarf und enthalten zudem keinerlei proprietäre Formatierungen. Jede Handschrift erhält eine zitierfähige Adressierung in Form einer persistenten URL und eines Uniform Ressource Name (URN). Die Metadaten können per OAI-Schnittstelle abgerufen werden und enthalten alle zur Nutzung durch den DFG-Viewer notwendigen Angaben. Mit der Archiv-Funktion des Workflow-Programms werden die Original-Scandateien zusammen mit den Metadaten im METS-XML-Format in ein separates Verzeichnis verschoben und auf Plattensystemen des Universitätsrechenzentrums Heidelberg archiviert. Die Ablieferung eines

Digital Master an die Deutsche Nationalbibliothek ist vorgesehen und soll umgesetzt werden, sobald Systeme für den Routinebetrieb zur Verfügung stehen (KOPAL).

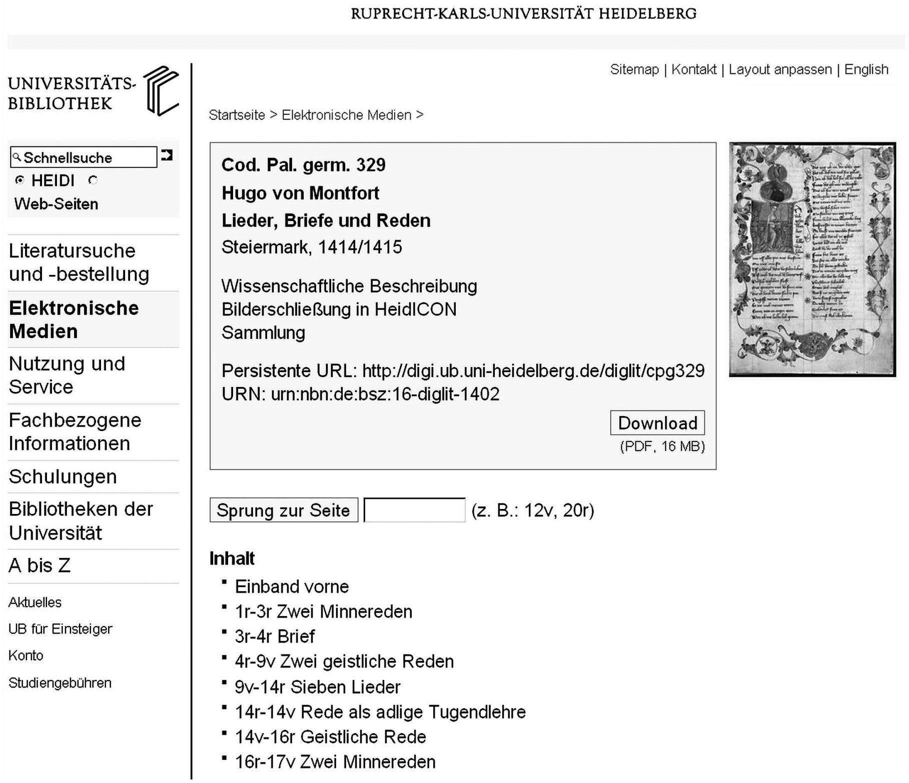
3 Präsentation der Handschriften und Benutzungsfunktionen

Die Bereitstellung der Daten im Internet ermöglicht dem Benutzer nicht nur einen orts- und zeitunabhängigen Einblick in die Handschriften, darüber hinaus werden erhebliche Arbeitserleichterungen geboten. Innerhalb des Webauftritts der UB Heidelberg erhält der Besucher freien Zutritt zu den »digitalen Bücherregalen« der Bibliotheca Palatina, aus denen er einzelne Bücher zur genaueren Betrachtung »herausnehmen« kann. Die Auswahl wird durch die nach Signaturen der Codices geordnete Übersicht sowie durch eine kurze inhaltliche Benennung und eine exemplarische Text- oder Bildseite, die als bildhafter Repräsentant des Codex dient, geleitet. Hinter dem Link der Handschrift selbst liegt die Bildschirmpräsentation des gesamten Buches, die es ermöglicht, eine beliebige Seiten- bzw. Blattzahl direkt anzusteuern, an den Anfang oder das Ende des Dokuments zu springen, oder auch seitenweise vor- bzw. zurückzublättern. Zusätzlich ist jede digitale Reproduktion mit weiteren Informationen und Navigationsmöglichkeiten angereichert: Die Werkeinstiegsseite enthält bibliographische Informationen wie Signatur, Autor, Titel, Herstellungsort und Datierung sowie ein Inhaltsverzeichnis mit einzeln anwählbaren Kapitelüberschriften (vgl. Abb. 2).

Letzteres wird auf der Grundlage der wissenschaftlichen Erschließung der Handschrift erstellt. Über eine »Vorschau«-Funktion kann sich der Benutzer zudem mithilfe von Thumbnails einen Überblick über die ganze Handschrift verschaffen. Von jeder einzelnen Seite (vgl. Abb. 3) können schließlich auch Arbeitskopien in verschiedenen Größen hergestellt werden, was bei schwer zu lesenden Texten oder bei Detailanalysen der Buchmalerei von Vorteil sein kann. Zudem kann jede Handschrift vollständig im PDF-Format heruntergeladen werden.

Den Digitalisaten unmittelbar beigefügt sind Links, die weiter führende Informationen bieten. Dabei handelt es sich zum einen um die wissenschaftliche Beschreibung der Handschrift, zum zweiten um den Zugang zu einer Datenbank, die das Bildmaterial der jeweiligen Handschrift erschließt und zum dritten um weiter führende Informationen zu den medizinischen Rezeptsammlungen.

Zu den ersten beiden Links gelangt der Benutzer über die Werkeingangsseite. Die im Rahmen der Rekatalogisierung der Codices Palatini germanici bereits erarbeiteten Handschriftenbeschreibungen (Kataloge Heidelberg, UB, Bd. 6-8) sind als PDF-Datei dem jeweiligen Codex beigefügt. So kann die gemäß den von der DFG vorgegebenen »Richtlinien Handschriftenkatalogisierung« vorgenommene wissenschaftliche Tiefenerschließung etwa zum Aufbau der Codices, zum Schreiber, zur Provenienz oder zum Bildschmuck direkt nachvollzogen werden. Für Codices, deren Katalogeinträge im Rah-


RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG
UNIVERSITÄTS-BIBLIOTHEK
Sitemap | Kontakt | Layout anpassen | English
Schnellsuche
● HEIDI ○ Web-Seiten
Literatursuche und -bestellung
Elektronische Medien
Nutzung und Service
Fachbezogene Informationen
Schulungen
Bibliotheken der Universität
A bis Z
Aktuelles
UB für Einsteiger
Konto
Studiengebühren
Startseite > Elektronische Medien >
Download
Sprung zur Seite

Cod. Pal. germ. 329
Hugo von Montfort
Lieder, Briefe und Reden
Steiermark, 1414/1415

Wissenschaftliche Beschreibung
Bilderschließung in HeidICON
Sammlung

Persistente URL: http://digi.ub.uni-heidelberg.de/diglit/cpg329
URN: urn:nbn:de:bsz:16-diglit-1402

(PDF, 16 MB)

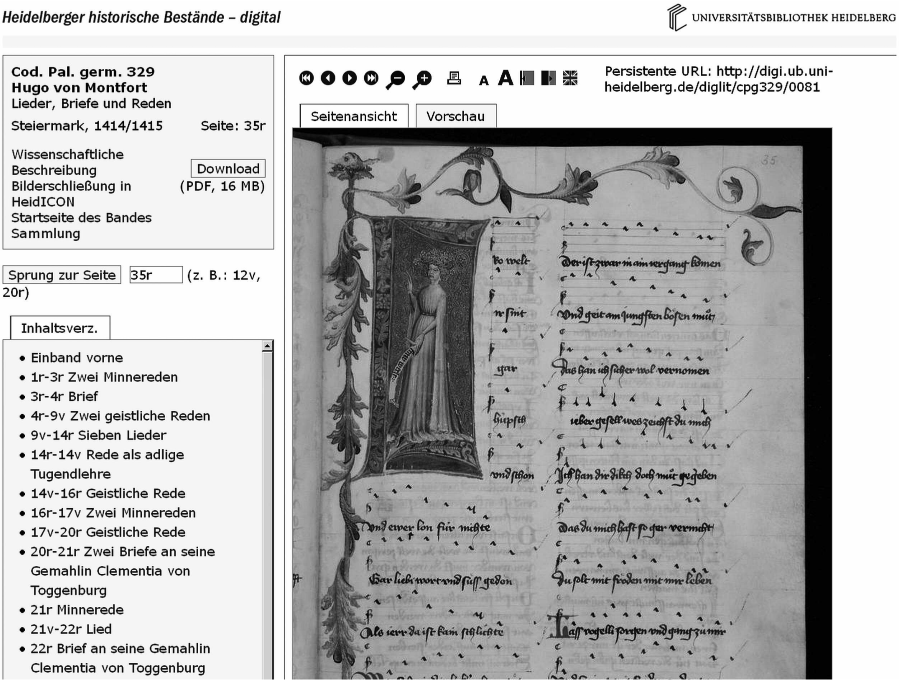
(z. B.: 12v, 20r)

Inhalt

- Einband vorne
- 1r-3r Zwei Minnereden
- 3r-4r Brief
- 4r-9v Zwei geistliche Reden
- 9v-14r Sieben Lieder
- 14r-14v Rede als adlige Tugendlehre
- 14v-16r Geistliche Rede
- 16r-17v Zwei Minnereden

Abbildung 2. Cod. Pal. germ. 329, Werkeinstiegsseite mit bibliographischen Informationen und Inhaltsverzeichnis.

men des laufenden Katalogisierungsprojektes noch nicht druckfertig erarbeitet sind, wird der Benutzer vorübergehend auf die älteren Handschriftenkataloge (Bartsch; Wille) verwiesen. Gleichfalls über die Werkeingangsseite kann der Benutzer auf die Bilddatenbank HeidICON zugreifen. Für das Projekt »Bibliotheca Palatina – digital« wurde dort ein eigener Pool für das Bildmaterial der Handschriften eingerichtet. Der in den wissenschaftlichen Beschreibungen in der Regel nur summarisch erfasste Buchmalereischmuck der Handschriften wird hier detailliert beschrieben und recherchierbar gemacht. HeidICON bietet differenzierte Verwaltungs- und Suchfunktionen, die einen individuellen Umgang mit dem Bildmaterial ermöglichen. Um die Auffindbarkeit von Bildern zu erhöhen, wird bei einigen Erschließungskategorien das Vokabular der Schlagwortnormdatei (SWD), das deutschlandweit in Bibliotheken auch für die Litera-



Heidelberger historische Bestände – digital

UNIVERSITÄTSBIBLIOTHEK HEIDELBERG

Cod. Pal. germ. 329
Hugo von Montfort
Lieder, Briefe und Reden
Steiermark, 1414/1415 Seite: 35r

Wissenschaftliche Beschreibung
Bilderschließung in HeidICON
Startseite des Bandes
Sammlung

Download (PDF, 16 MB)

Sprung zur Seite 35r (z. B.: 12v, 20r)

Persistente URL: http://digi.ub.uni-heidelberg.de/diglit/cpg329/0081

Inhaltsverz.

- Einband vorne
- 1r-3r Zwei Minnereden
- 3r-4r Brief
- 4r-9v Zwei geistliche Reden
- 9v-14r Sieben Lieder
- 14r-14v Rede als adlige Tugendlehre
- 14v-16r Geistliche Rede
- 16r-17v Zwei Minnereden
- 17v-20r Geistliche Rede
- 20r-21r Zwei Briefe an seine Gemahlin Clementia von Toggenburg
- 21r Minnerede
- 21v-22r Lied
- 22r Brief an seine Gemahlin Clementia von Toggenburg

Abbildung 3. Cod. Pal. germ. 329, fol. 35r.

turerschließung eingesetzt wird, benutzt. Die eigentliche Bilderschließung konzentriert sich im Wesentlichen auf die ikonographische Bestimmung von Einzeldarstellungen. Bei Buchgattungen mit festgelegtem Motivrepertoire, wie beispielsweise dem »Speculum humanae salvationis« oder der »Biblia pauperum«, sind in der Datenbank lediglich das Bildthema und etwaige auffällige Abweichungen kurz notiert. Bei Buchgattungen hingegen, die narrative, den Text begleitende Bildzyklen enthalten, basiert die Bildbeschreibung auf einer Analyse im Verhältnis zu den Texten.

Für den Bereich der medizinischen Rezeptsammlungen wurden im Kontext der Rekatalogisierung zu jeder einzelnen Handschrift Katalogisate erarbeitet, in die alle Rezeptüberschriften aufgenommen sind, während im gedruckten Katalog aus Platzgründen nur eine summarische Darstellung der Gliederung vorgenommen werden konnte. Diese »Langversionen« der Handschriftenkatalogisate sind in mehreren PDF-Dateien zusammengefasst. Ihre Bereitstellung ermöglicht es, in dem bisher kaum erforschten Bereich der medizinischen Handschriften insbesondere des 16. Jahrhunderts die Genese der Texte nachzuvollziehen. Dies ist zum einen für die Erforschung der Zusammenhänge innerhalb der zahlreichen medizinischen Handschriften der Bibliotheca Palatina, zum anderen aber auch für eine künftige Untersuchung im größeren Zusammenhang ein hilfreiches Instrument.

Der allgemeine Zugang zur digitalen Bibliothek gibt der Arbeit mit den mittelalterlichen Überlieferungsträgern ein völlig neues Gesicht. Da die wertvollen Unikate nicht mehr unmittelbar in die Hand genommen werden müssen und die Kosten für die Reproduktion entfallen, ist eine Arbeit mit den Materialien insbesondere im akademischen Unterricht erleichtert. Die Einsichtnahme in die Texte und Bildmaterialien ist jedermann jederzeit möglich, zudem wurden die digitalen Sekundärformen der Handschriften auch im Südwestdeutschen Bibliotheksverbund (SWB) verzeichnet. Sie sind so – gemeinsam mit dem Druckschriftenbestand – auch im Heidelberger Online-Katalog HEIDI recherchierbar. Dies kann selbstverständlich nicht in Form einer stark differenzierten Erschließung und Beschreibung geschehen, wie sie für die gedruckten Kataloge durch die DFG-Richtlinien (Richtlinien Handschriftenkatalogisierung) vorgeschrieben ist, sondern erfolgt in Form von Kurzaufnahmen. Diese ermöglichen zum einen die eindeutige Identifizierung der Handschrift und bieten direkten Zugriff auf die digitalisierte Handschrift selbst sowie auf die dort verknüpfte, ausführliche wissenschaftliche Beschreibung (Zur Aufnahme von Handschriften in Online-Kataloge vgl. Fabian u.a.). Darüber hinaus betreibt die UB Heidelberg auch die Erfassung der Handschriften in dem nationalen Nachweisinstrument für Handschriften »Manuscripta medievalia«. Auch im deutschen MICHAEL-Portal, das bundesweit digitale Sammlungen und Bestände aus Archiven, Bibliotheken und Museen listet und zentral zugänglich macht, sind die Heidelberger digitalen Sammlungen enthalten. Durch die Beteiligung des deutschen MICHAEL-Portals am multilingualen europäischen MICHAEL-Portal (Multilingual Inventory of Cultural Heritage in Europe) wird darüber hinaus das digitale Heidelberger Kulturgut auch für ein europäisches und weltweites Publikum besser verfügbar. Gleiches gilt für die Kooperation mit dem europäischen Projekt ENRICH (European Networking Resources and Information concerning Cultural Heritage). Aufbauend auf der Datenbank »Manuscriptorium« wird hier zusammen mit zahlreichen internationalen Partnern eine europäische digitale Handschriften-Bibliothek aufgebaut.

4 Nutzung der Digitalisate

Die Digitalisierung eröffnet zahlreiche neue Möglichkeiten im Umgang mit den Handschriften. Insbesondere im Bereich der Lehre ist eine Heranführung an die mittelalterlichen und frühneuzeitlichen Überlieferungsträger deutlich erleichtert. Konnte man sich bisher in der Regel nur mit Hilfe der gedruckten Kataloge oder im Einzelfall unter Hinzuziehung von Faksimiles über einzelne Handschriften informieren, so ist es nun möglich, die Informationen, die das Katalogisat bietet, unmittelbar am Bildschirm nachzuvollziehen. Die Beschreibungsprinzipien, die sich dem ungeübten Studierenden meist nicht auf Anhieb erschließen, können Schritt für Schritt nachvollzogen werden. Was bisher vor Ort bei Handschriftenexkursionen nur in einem sehr eng gesteckten Zeitrahmen möglich war, kann nun ausgiebig von zu Hause aus vorbereitet werden. Da es einfach möglich ist, Textpassagen zu vervielfältigen, kann auch das Lesen handschriftlicher Texte an den verschiedensten Proben eingeübt werden. Dass diese Möglichkeiten auch genutzt werden, spiegelt sich in den Nutzungszahlen der vergangenen Monate wider,[2] von denen sich einige im Zusammenhang mit Besuchen von Gruppen in der Heidelberger Handschriftenabteilung lesen. So wurde der Cod. Pal. germ. 346, der eine bebilderte Version von Eilharts von Oberge »Tristrant« enthält, im Rahmen einer Seminarveranstaltung behandelt und in den untersuchten Monaten 20 Mal pro Monat abgerufen. Im Rahmen derselben Veranstaltung wurden auch die Cod. Pal. germ. 60 und 349 betrachtet und weisen Nutzungszahlen von durchschnittlich 22 bzw. 8 Zugriffen pro Monat auf. Der Cod. Pal. germ. 430, enthaltend Hans Leckúchners » Kunst des Messerfechtens«, wurde im Zusammenhang mit einer anderen Lehrveranstaltung, der eine Exkursion nach Heidelberg folgte, im untersuchten Zeitraum 50 Mal pro Monat angewählt. Inwieweit das Angebot in der akademischen Lehre genutzt wird, ohne dass ein Besuch vor Ort die Veranstaltung ergänzt, kann nur vermutet werden. Zugriffszahlen von 80 pro Monat, wie sie in der Zeit vom November 2008 bis Februar 2009 Heinrichs von Veldeke »Eneas« (Cod. Pal. germ. 403) aufweist, oder 55 pro Monat beim »Rolandslied« des Pfaffen Konrad (Cod. Pal. germ. 112) sowie 30 pro Monat bei Ulrichs von Zazikhofen »Lanzelet« (Cod. Pal. germ. 371) lassen jedenfalls auf verschiedene, gemeinsam an diesem Text arbeitende Einzelnutzer schließen.

Bebilderte Handschriften hatten schon immer eine höhere Attraktivität als solche, die reinen Text enthalten; dies zeigt sich auch an der Auswahl von Handschriften aus dem digitalen Bücherregal. Die höchsten Zugriffszahlen weisen Handschriften mit reichhaltigem und interessantem Bildmaterial auf: Der Codex Manesse (Cod. Pal. germ. 848) wurde pro Monat durchschnittlich 2541 Mal aufgeschlagen, der Heidelberger Sachsenspiegel (Cod. Pal. germ. 164) 274 Mal, die bebilderte Parzivalhandschrift Cod. Pal. germ. 339 über 200 Mal, die bebilderte Ackermannhandschrift (Cod. Pal. germ. 76) 81 Mal. Erst an sechster Stelle findet sich mit der »Kleinen Heidelberger Liederhandschrift« (Cod. Pal. germ. 357) ein Objekt, das kein Bildmaterial enthält. Wissenschaftlich genutzt wird das Bildmaterial z. B. in dem kunsthistorischen Projekt »Stadt im Bild: Die Ausformung

[2] Für die Monate November 2008 bis Februar 2009 wurde eine erste Zugriffsstatistik erhoben. Sie gibt Aufschluss darüber, wie viele Zugriffe pro Tag und Rechner über die Eingangsseite des Projektes auf einzelne Handschriften vorgenommen wurden. Zugriffe über Suchmaschinen wurden ignoriert. Nicht berücksichtigt werden konnte leider auch der wichtige Fall, dass ganze Handschriften als PDF auf den eigenen PC heruntergeladen werden, ein Vorgehen, welches bei längerfristiger regelmäßiger Nutzung einer Handschrift der Regelfall sein dürfte. Die Statistik ist wegen des kurzen Beobachtungszeitraums von begrenzter Aussagekraft, kann aber zumindest erste Anhaltspunkte über die tatsächliche Nutzung geben.

eines städtischen Selbstbildes in der Augsburger Buchillustration zwischen Spätmittelalter und Früher Neuzeit«, das in Heidelberg am Lehrstuhl von Prof. Dr. Liselotte E. Saurma angesiedelt ist. Die Sichtung von Bildmaterial ist aber auch im Bereich der Ausstellungskonzeption ein wichtiger Aspekt. So wurden die Digitalisate beispielsweise bei der Vorbereitung der Ausstellung »Rituale und die Ordnung der Welt« verwendet, die vom Oktober 2008 bis Ende Januar 2009 in den Ausstellungsräumen der Universitätsbibliothek Heidelberg zu sehen war. Eine große Zahl möglicher Exponate ist durch das Digitalisat rasch einsehbar. Vorentscheidungen für die Auswahl konnten so bequem vom eigenen Schreibtisch aus getroffen werden, ohne dass bereits in diesem Schritt die Originale strapaziert werden mussten.

Auch für die Arbeit mit dem Textmaterial ist das Vorliegen der Digitalisate von großem Nutzen. Dies wird an einem Beispiel aus der Tätigkeit im Rahmen der Rekatalogisierung der Codices an der UB Heidelberg deutlich: Unter den Handschriften befinden sich zahlreiche medizinische Rezeptsammlungen, deren Erschließung eine wissenschaftliche Pionieraufgabe ist. Um das Material leichter strukturieren zu können, wird zunächst eine ausführliche Transkription der Rezeptüberschriften erstellt. Diese können bequem am Rechner angelegt werden, wobei das digitalisierte Original und die Transkription direkt nebeneinander auf einem Bildschirm aufgeschlagen werden. Die Handschrift selbst wird hierfür zunächst nicht benötigt; sie wird erst für den Korrekturdurchgang aufgeschlagen. Darüber hinaus ermöglicht es die digitale Bibliothek, relativ rasch Zusammenhänge zwischen verschiedenen Rezeptsammlungen zu eruieren: Über die als PDF einsehbaren Langversionen sind Rezeptzuträger und Rezepttitel ermittelbar. Findet sich eine mögliche Parallele, ist es ohne größeren Aufwand möglich, die Stelle direkt in der Online-Version des Buches nachzuschlagen und einzelne Textpassagen im Detail zu vergleichen. Vermutungen über Vorlagen können so schnell erhärtet oder widerlegt werden.

Gerade das Nachprüfen einzelner Details ist ein großer Vorteil der Digitalisierung. Selbst wenn die Originalmaterialien vor Ort vorhanden sind oder als Mikrofilme zur Verfügung stehen, ist das Nachschlagen mit Hilfe des Digitalisates einfacher und schneller zu bewerkstelligen. Wichtiger noch ist dieser Aspekt bei Forschern, die nicht vor Ort tätig sind. Nicht immer ist es möglich und verhältnismäßig für einzelne Fragen die Handschrift selbst einzusehen. Musste man sich bisher auf die Expertise der Mitarbeiter vor Ort verlassen, kann nun der Forscher von seinem heimischen Schreibtisch aus in die Handschrift schauen und sich bei der Entwicklung seiner weiteren Fragen von dem leiten lassen, was der Text an zusätzlichen Informationen bereithält, nach denen er vielleicht ansonsten gar nicht gefragt hätte.

Ähnlich sieht es im Bereich der Schreiberhände aus. Wenn Ähnlichkeiten zwischen Schreibern vermutet werden, ist es rasch möglich, diese Vermutung zu prüfen. Einzelne Hände können nebeneinandergelegt oder direkt ausgedruckt und verglichen werden. Im Bereich der digitalen Bilderkennung wird die Heidelberger Liederhandschrift Hugos von Montfort (Cod. Pal. germ. 329) zudem für ein Pilotprojekt an der Universität Graz benutzt, das Projekt »Datenbank zur Authentifizierung mittelalterlicher Schreiberhände« (DAmalS).[3]

Die Digitalisierung der Handschriften entzaubert in gewisser Weise den Umgang mit mittelalterlichem Material. Die Einsichtnahme in die handschriftliche Überlieferung ist nicht mehr der Heilige Gral, der nur Auserwählten zu bestimmten Zeiten und nach vorheriger Planung und Anreise zugänglich ist. Gleichzeitig ermöglicht die Technik es, in der Öffentlichkeit ein breiteres Verständnis für die Bedeutung der alten Kulturgüter zu wecken und zu erhalten. Dabei wird die Arbeit mit dem Buch allerdings keineswegs überflüssig. Vielmehr kann man sich als Forscher nun auf die Arbeit mit dem Buch optimal vorbereiten und sich in der kurzen Zeit, die man mit dem Buch vor Ort verbringen kann, auf solche Fragen konzentrieren, für die tatsächlich das Original benötigt wird.

Bibliographie

Bartsch, Karl. *Die altdeutschen Handschriften der Universitäts-Bibliothek in Heidelberg.* Heidelberg: Koester, 1887.

Bibliotheca Palatina Digital. <http://palatina-digital.uni-hd.de>.

DFG-Viewer. <http://dfg-viewer.de>.

Doerr, Wilhelm u.a., Hrsg. *Semper Apertus. Sechshundert Jahre Ruprecht-Karls-Universität Heidelberg 1386–1986.* Festschrift in sechs Bänden. Band I: Mittelalter und Frühe Neuzeit 1386–1803. Berlin: Springer, 1985.

Effinger, Maria, Margit Krenn und Thomas Wolf. »Der Vergangenheit eine Zukunft schaffen: Die Digitalisierung der Bibliotheca Palatina in der Universitätsbibliothek Heidelberg.« *B.I.T. online*, 11.2 (2008): 157–166.

Empfehlungen der DBV AG Handschriften/Alte Drucke zur Herstellung, Internetpräsentation und Verwaltung von Digitalisaten alter Drucke und Handschriften. Berlin: Bibliotheksverband, 2006. <http://www.bibliotheksverband.de/aghandschriften/dokumente/digi-empfehlungen.html>.

Fabian, Claudia, Wolfgang-Valentin Ikas und Mathias Kratzer. »Vom Nutzen der Vernetzung und den Chancen der Digitalisierung: neue Wege der Handschriftenerschließung in der Bayerischen Staatsbibliothek.« *Zeitschrift für Bibliothekswesen und Bibliographie* 54 (2007): 322–335.

Fuchs, Christoph. *Dives, pauper, nobilis. Magister, frater, clericus. Sozialgeschichtliche Untersuchungen über Heidelberger Universitätsbesucher des Spätmittelalters (1386–1450).* Leiden u.a.: Brill, 1995.

Jammers, Ewald. »Zur Geschichte der Heidelberger Universitäts-Bibliothek und ihrer Quellen.« *Ruperto-Carola, Sonderband. Aus der Geschichte der Universität Heidel-*

[3] Siehe den Beitrag von Hofmeister u.a. in diesem Band.

berg und ihrer Fakultäten. Aus Anlaß des 575jährigen Bestehens der Ruprecht-Karl-Universität Heidelberg Hrsg. Gerhard Hinz. Heidelberg: Brausdruck, 1961. 112–133.

Kataloge Heidelberg, UB, Band 6–8 und 10:

– *Die Codices Palatini germanici in der Universitätsbibliothek Heidelberg (Cod. Pal. germ. 1–181).* Bearb. von Karin Zimmermann unter Mitwirkung von Sonja Glauch, Matthias Miller und Armin Schlechter. Wiesbaden: Reichert, 2003 (Kataloge der Universitätsbibliothek Heidelberg 6).

– *Die Codices Palatini germanici in der Universitätsbibliothek Heidelberg (Cod. Pal. germ. 182–303).* Bearb. von Matthias Miller und Karin Zimmermann. Wiesbaden: Harrassowitz, 2005 (Kataloge der Universitätsbibliothek Heidelberg 7).

– *Die Codices Palatini germanici in der Universitätsbibliothek Heidelberg (Cod. Pal. germ. 304–495).* Bearb. von Matthias Miller und Karin Zimmermann. Wiesbaden: Harrassowitz, 2006 (Kataloge der Universitätsbibliothek Heidelberg 8).

– *Die Codices Palatini germanici in der Universitätsbibliothek Heidelberg (Cod. Pal. germ. 496–659).* Bearb. von Pamela Kalning, Matthias Miller und Karin Zimmermann (Kataloge der Universitätsbibliothek Heidelberg 10) (in Vorbereitung).

Manfred-Lautenschläger-Stiftung. <http://www.manfred-lautenschlaeger-stiftung.de>.

Manuscripta Mediaevalia. <http://www.manuscripta-mediaevalia.de>.

Manuscriptorium. <http://www.manuscriptorium.eu>.

Michael. Multilingual Inventory of Cultural Heritage in Europe. <http://www.michael-culture.org>.

MichaelDeutschland. Multilingual Inventory of Cultural Heritage in Europe. <http://www.michael-portal.de>.

Miethke, Jürgen. *Libri actorum Universitatis Heidelbergensis, Series A, Acta Universitatis Heidelbergensis, Bd 1, 1386–1410 (zugleich das erste Amtsbuch der Juristischen Fakultät).* Hrsg. Jürgen Miethke, bearbeitet von Heiner Lutzmann und Hermann Weisert. Heidelberg: Winter, 1986.

Mittler, Elmar. »Die Bibliotheca Palatina. Skizzen zu ihrer Geschichte.« *Mit der Zeit. Die Kurfürsten von der Pfalz und die Heidelberger Handschriften der Bibliotheca Palatina.* Hrsg. Elmar Mittler und Wilfried Werner. Wiesbaden: Reichert, 1986: 7–50.

MODS. *Metadata Object Description Schema.* Washington (DC): Library of Congress, 2009. <http://www.loc.gov/standards/mods>.

Moraw, Peter. »Heidelberg: Universität, Hof und Stadt im ausgehenden Mittelalter.« *Studien zum städtischen Bildungswesen des späten Mittelalters und der frühen Neuzeit.* Hrsg. Bernd Moeller u.a. Göttingen: Vandenhoeck & Ruprecht, 1983: 524–552.

Praxisregeln im Förderprogramm »Kulturelle Überlieferung.« Deutsche Forschungsgemeinschaft. Bonn: DFG, 2008.
<http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/aktuelles/download/praxisregeln_kulturelle_ueberlieferung_0208.pdf>.

Richtlinien Handschriftenkatalogisierung. Deutsche Forschungsgemeinschaft, Unterausschuß für Handschriftenkatalogisierung. 5. erw. Auflage. Bonn-Bad Godesberg: Deutsche Forschungsgemeinschaft, 1992.

Ritter, Gerhard, Hrsg. *Die Heidelberger Universität. Ein Stück deutscher Geschichte. Erster Band: Das Mittelalter (1386–1508).* Heidelberg: Winter, 1936.

Stadt im Bild: Die Ausformung eines städtischen Selbstbildes in der Augsburger Buchillustration zwischen Spätmittelalter und Früher Neuzeit. DFG-Projekt am Zentrum für Europäische Geschichts- und Kulturwissenschaften. Heidelberg: Institut für Europäische Kunstgeschichte, 2008.
<http://www.khi.uni-heidelberg.de/projekte/alteabt/augsburg.htm>.

Wilken, Friedrich. *Geschichte der Bildung, Beraubung und Vernichtung der alten Heidelbergischen Büchersammlungen. Nebst einem Verzeichniß der im Jahr 1816 von dem Pabst Pius VII. der Universität Heidelberg zurückgegebenen Handschriften und einigen Schriftproben.* Heidelberg: Oswald, 1817.

Wille, Jakob. *Die deutschen Pfälzer Handschriften des XVI. und XVII. Jahrhunderts der Universitäts-Bibliothek in Heidelberg. Mit einem Anhange: Die Handschriften der Batt'schen Bibliothek.* Verzeichnet und beschrieben von Jakob Wille. Heidelberg: Koester, 1903.

Manuscriptorium Digital Library and ENRICH Project: Means for Dealing with Digital Codicology and Palaeography

Zdeněk Uhlíř, Adolf Knoll

Abstract

Codicology and palaeography in the digital age can be developed both through adapting existing methods and using information and communication technologies. This can be achieved e.g by projects focusing on the integration of individual resources under a single user interface. This is the aim of the Manuscriptorium digital library as well as the ENRICH project. The integration is based on the centralisation of metadata from various resources and on the distributed storage of data, mainly digital images. This is implemented through a distributed complex digital document, containing the so-called identification record and more data types. The construction of the integrated Manuscriptorium digital library within the ENRICH project is being done in four basic ways: automatically, or semi-automatically respectively manually, and those both online and offline. This has made it possible to amass more than 5,000 documents. For Manuscriptorium, a search is important, which allows information to be gathered through special fields and the differences in graphics to be harmonised. The aim of the ENRICH project is also the creation of tools for the compilation of virtual collections and documents. In its method of integrating resources, the Manuscriptorium endeavours to be an instrument of codicological and palaeographic research.

Zusammenfassung

Kodikologie und Paläographie können sich im digitalen Zeitalter durch die Veränderung bestehender Methoden und durch die Nutzung von Informations- und Kommunikationstechniken weiterentwickeln. Diese Entwicklung kann beispielsweise durch Projekte befördert werden, die auf die Integration einzelner Ressourcen in ein gemeinsames Nutzerinterface abzielen. Dies ist das Ziel sowohl der Digitalen Bibliothek von Manuscriptorium als auch des ENRICH Projektes. Die Integration basiert dabei auf der Zentralisierung der beschreibenden Metadaten aus verschiedenen Sammlungen und auf der verteilten Datenhaltung der eigentlichen Daten, d.h. digitaler Bilder. Erreicht wird dies durch ein verteiltes komplexes Datendokument, welches einen sogenannten Identifizierungdatensatz und weitere Datentypen enthält. Der Aufbau der integrierten

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 67–78.

Digitalen Bibliothek Manuscriptorium im Rahmen des ENRICH Projektes geschieht auf vier Wegen: automatisch oder halbautomatisch bzw. händisch, und zwar beides entweder on- oder offline. So war es möglich, mehr als 5.000 Dokumente zusammenzuführen. Für Manuscriptorium ist eine Suche, die es ermöglicht, Information aufgrund von Spezialfeldern zu finden, ebenso wichtig wie der Versuch, die Unterschiede in der Darstellung zu harmonisieren. Ein Ziel des ENRICH Projekts ist die Entwicklung von Tools zur Zusammenstellung von virtuellen Sammlungen und Dokumenten. Durch diese Ressourcenintegration stellt Manuscriptorium ein Instrument für kodikologische und paläographische Forschung dar.

1 Introduction

Since the 1990s, people have been talking about the *third, i.e. information, wave* (Toffler), an *information and knowledge society*, and about the *digital economy.* Since the eve of the millennium, they have also been discussing *digital history* and/or *digital philology*, or digital research in general.

Hence, we can speak about *digital codicology* and/or *palaeography*, or about other historical auxiliary sciences (*historische Hilfswissenschaften*) in particular. In the digital age, codicology and palaeography can be developed in two directions: 1) by continuing to adapt the traditional methodologies that have been developed since the 16th century and 2) by updating information and communication technologies as an opportunity for an innovation of the methodology. Both directions can evolve from the resources that provide access to our written cultural heritage, especially to manuscript books and earlier administrative materials. This can be best achieved by special projects, which concerns not only specific-focus projects (dealing e.g. with watermarks, book covers, scripts, scribes), but also projects focused on the integration of particular resources into a uniform interface.

The development of codicological and palaeographic methodologies in the digital age is the main aim of the Manuscriptorium and ENRICH projects.

2 The Manuscriptorium and ENRICH

Both the Manuscriptorium and ENRICH[1] projects plan to become a component part of the European Digital Library for the specific cultural heritage consisting of manuscripts,

[1] The technical partners of ENRICH are as follows: AiP Beroun Ltd., Beroun, Czech Republic; Oxford University Computing Services, Oxford, United Kingdom; Media Integration and Communication Centre, Florence, Italy; Institute of Mathematics and Informatics, Vilnius, Lithuania; SYSTRAN S.A., Paris, France; Computer Science for the Humanities, Cologne, Germany; Poznan Supercomputing and Networking Center, Poznan, Poland.

incunabula, early printed books, historical archival materials, maps etc. that would be easily accessible and searchable using various sophisticated approaches as well as multilingual search and ontologies. When saying 'a *European* digital library', we do not mean a single digital library used all over Europe but such a resource that integrates and provides real European content. As a general goal, the Manuscriptorium and ENRICH projects want to accumulate a critical mass of the digitised manuscripts deposited in the memory institutions in European countries or of European origin. In other words, the Manuscriptorium and ENRICH projects aspire to be fundamental sources for manuscript study on a global scale. Accordingly, there are three main goals to be achieved in three steps.

2.1 Integration of Resources

The integration of resources is not simple—precisely on the contrary, it is a very complex task. As the first step, it supposes the accumulation of metadata, i.e. not merely descriptive metadata, that is manuscript catalogue records, but also other kinds of metadata, especially structural metadata, representing the physical structure of the original document (mostly a book-codex), as well as the creation of a compound digital document (the so-called virtual book). As the second step, it assumes the accumulation of data, i.e. usually digital images of manuscripts but also full electronic texts (editions of historical documents) and also electronic representations of music notation or even audio or video files. The identification of documents (both physical manuscripts and their digital copies) is substantial, because it enables any orientation and navigation within the whole digital library and any kind of heuristics within the manuscript collections represented in the digital library. The heart of the Manuscriptorium system is therefore the database of identification records. In other words, every digital copy of a manuscript and/or any full text needs a descriptive record; on the other hand, there may be descriptive records without any digital images and/or full text. Furthermore, one physical item may have multiple descriptive records within the Manuscriptorium system although only one of them is and can be an identification record, because it supports the search utility within the digital library. Consequently, an identification record is in correlation to both the original document and its digital copy such that there is only one original document, one digital copy and one identification record. Multiplex descriptive records referring to one physical item can be both language variants and/or alternate records, i.e. these records may be translations of the same record or created from the points of view of different research interests (e.g. palaeography, codicology, heraldry, literary history, art history, musicology, history of theology and/or philosophy, history of science, etc.). However, as long as the identification record is just one of them, it is an unambiguous record (perhaps better put as *ein-eindeutig* in German or *jedno-jednoznačný* in Czech).

The identification record and the database of identification records enable the preparation of a compound digital document, i.e. the creation of a virtual book. Since the Manuscriptorium system was launched in 2003, compound digital documents (virtual books) have been created according to the XML-based standard called MASTER+, which is an extension of the MASTER developed in 1999–2001 by a consortium led by the De Monfort University Leicester, in which also the National Library of the Czech Republic participated. In the course of the ENRICH project in 2007–2009, the chapter on the manuscript description and the TEI P5 schema for ENRICH was defined, therefore any document in conformance with the ENRICH schema is also a TEI-conformant document and can be used by any TEI-aware software. This is very important, because when the ENRICH project comes to an end, the Manuscriptorium system will be widely interoperable and that interoperability is a significant step towards the easy integration of new resources.

Interoperability is naturally essential for imports/uploads and/or exports/downloads of metadata/data. However, it is not only a means for building centralised resources but also the basic condition for creating distributed resources—and work within the virtual distributed digital environment is what the Manuscriptorium is all about. A distributed compound digital document in general means that its parts are distributed anywhere throughout cyberspace. A distributed compound digital document as it is understood for the Manuscriptorium's purposes means in particular that the metadata part is centralised in the database of the identification records of the Manuscriptorium. On the other hand, the data remains at the server of the content partner. That is why not only the descriptive but also the structural metadata are of equal importance; the creation of the distributed compound digital document is impossible without the structural metadata, which mark the subordinated items (usually pages) of the appropriate physical document (usually codex), thus enabling the creation of a virtual book and its browsing as well as searching (when using so-called *applied foliation/pagination* as the labelling/names for individual files). This is of course not enough, because the necessary condition for the successful integration of documents distributed in cyberspace is that the data referred to by hyperlinks from the centralised structural metadata resource must be stored in a stable and persistent repository with perpetual and unmistakable names. It is a *conditio sine qua non* for the creation of any operating distributed resource.

Thus, the persistent addressing is a fundamental issue. The question may be whether persistent identification is possible when understood absolutely, i.e. regardless of a concrete resource and/or information system. On the other hand, such an absolutely understood question (i.e. the question of a universal persistence) has little sense—if any—for the practical process of building a digital manuscript library, although it has a deep meaning in theory. Thus, only a question of a relative (system-related) persistence, i.e. the solution of persistent addressing and identification regarding the Manuscriptorium system, is connected with a topic that concerns digital palaeography and codicology. The way of creating persistent identification within the Manuscriptorium is quite simple, consisting of:

- The identification of the owner of the physical object (that relates to the appropriate digital object, of course), the identification of the physical object (shelf mark, call number etc.);
- The identification of a part of the physical object (typically a page, but also an open book, i.e. the back page of a preceding leaf and the front page of the next leaf altogether);
- The identification of the quality level of the images (e.g. gallery thumbnail, preview, normal size, black and white optimisation, etc.).

Such a kind of persistent addressing makes it possible both to ascertain and find an individual digital object (i.e. page) as well as to create an appropriate sequence of objects (i.e. sequence of images of the appropriate quality level) or complete the whole compound digital document independently of its centralised or distributed form. Thus, the use of consistently and logically created names for individual digital objects (files, not just documents) is crucial.

So as to ensure the smooth operation and progress of the Manuscriptorium system, still another condition regarding metadata is necessary. The Manuscriptorium has numerous partners, both current and potential, most of whom do not create their metadata solely for the Manuscriptorium system but for various, even divergent, goals and purposes, therefore using diverse metadata standards.

For the partners' results, i.e. data outputs, to be integrated, into the Manuscriptorium internal format (at present this is still MASTER+, but in the near future within the ENRICH implementation of TEI P5 we will use enrich.dtd, and enrich.xsd and the ENRICH Relax NG Scheme), metadata mapping and metadata conversion are needed. It is not an obsessive idea to transform heterogeneity into homogeneity but a practical effort to harmonise different metadata formats for the purpose of indexing them, which is the basis for efficient, if not for any, search. A digital library without efficient search (with no search or with mere browsing utilities) is not advisable for scholars or any other refined use.

On the basis of a deeper analysis of the particular content resources available within the frame of ENRICH, there are four typical ways of cooperation in integrating resources and in creating distributed compound digital documents:

1. The M-Tool (creating structural and descriptive metadata for individual documents): the M-Tool is an application serving the partners of the Manuscriptorium who create new metadata for existing image data. It enables the creation of descriptive and structural metadata and consequently also compound digital documents. The M-Tool currently exists as a stand-alone application and is going to be created

in an online version. It is now (end of April 2009) being tested. In its online version, it has been adapted to the new demands in the area of produced structural and descriptive metadata and to the usage of UTF-8.

2. The offline automated generation of structural metadata and connector for existing descriptive metadata: this method is similar to the previous one and suitable for partners who have a greater number of catalogue records which justify the creation of connectors. The use of a connector will make it possible to connect with the partner's source of primary information directly and will not necessitate the creation of duplicate descriptive metadata using the M-Tool. The M-Tool will be used to create structural metadata and descriptions in identification records, with the detailed description being processed by the connector. In the Manuscriptorium, it is possible to supply the outputs of the M-Tool with these brief descriptions through identifiers. These identifiers enable the document to be joined with the descriptive information processed by the connector, which is going to be optimised and adapted to the properties of input metadata to provide, where possible, zero-loss conversions.
3. The offline connector (converting the existing structural and descriptive metadata content offline): this is a tool enabling the conversion of the existing input metadata of compound digital documents depending on the needs of the Manuscriptorium system. It is part of the input interface of this system and is constructed in such a way that it cannot overreach the managing system of a partner. This approach makes it possible to separate the data production and management from the data presentation as well as optimise both of these tasks. It also ensures that it is not necessary to modify the existing processes at the workplaces of individual content partners. The most important parts of the connector are the conversion routines for transferring descriptive and structural metadata. This is suitable for partners who manage existing structural and descriptive metadata for such a number of documents that justifies the creation of the connector.
4. The online connector (converting the existing structural and descriptive metadata content using the OAI-PMH interface): it is similar to the offline connector but uses OAI-PMH for communication. At present, all the partners which have the OAI interface implemented will contribute with a relevant amount of documents justifying the creation of connectors.

In the course of its activity (since 2003), the Manuscriptorium team has developed a network of almost 50 Czech partners and more than 40 foreign partners. Some of them are quite small memory institutions (libraries, archives, museums etc.), yet others are the biggest and the most eminent players in manuscript digitisation (e.g. the National Library of the Czech Republic and the Czech national digitisation programme *Memoriae mundi series Bohemica*, the University Library Wrocław in Poland, the National and

University Library of Iceland and the Arne Magnusson Institute in Reykjavík in Iceland, the University of Cologne and the University Library Heidelberg in Germany, the Central National Library Florence in Italy, the National Library of Serbia, the National Library of Romania, the National Library of Spain etc.). In addition, the Manuscriptorium team and the partners from the ENRICH consortium negotiate with other important partners (e.g. the National Library of Austria, the National Library of Belarus, the State Library in Moscow and the Monastery Library at Sergiev Posad in Russia, the Badische Landesbibliothek Karlsruhe in Germany etc.).

At present (the end of April 2009), the Manuscriptorium digital library includes:

- 185,719 descriptive catalogue records for manuscripts, incunabula, early printed books, historical maps etc.;
- 5,133 compound digital documents (digital copies of manuscripts, incunabula, early printed books, historical maps etc.);
- Almost 1,500,000 individual images (typically pages of manuscripts, incunabula, early printed books), usually in five quality levels (gallery thumbnail, preview, low, normal/excellent, black and white optimisation);
- 350 compound digital documents containing full texts (transcriptions and/or editions of original historical documents).

2.2 Sophisticated Search

On the whole, search is the most important access method for the end user. Browsing is not enough, although several institutions that provide digitised manuscripts offer only this possibility of orientation and navigation within the resource. The possibility of search reduced to the mere library and shelf mark or call number is also unsatisfactory. Search applicable in any large digital library has to be based on more sophisticated prerequisites that use indexing, exact phrases, model sentences etc. Consequently, both a full text search and a search according to selected individual fields/elements should be implemented including the possibilities of a simple, combined and expert search with filters. The question is which fields/elements should be indexed for searching. Such a question is not trivial, because there are many differing aspects to consider. The following possibilities are implemented in the Manuscriptorium: as default it is search by Settlement and Repository and as optional it is by Country, Shelf Mark, Alternative Name, Title, Author, Place of Origin, Date of Origin, Name, Date, Rubric, Incipit, Explicit, Colophon, Responsibility, Responsibility-Name, Scribe, Additions, Origin, Provenance, Bibliography, Printer (the Manuscriptorium digital library provides incunabula and early printed books as well) or Music Notation. Of course, another selection of fields/elements could be discussed; however, the existing one is sufficient according to the feedback of our end users.

When integrating resources, there is yet another challenge for the Manuscriptorium digital library as an aggregator: discrepancies in manuscript descriptions. Because of heterogeneity of standards and of the different levels/depth of manuscript descriptions, the accessible data for machine processing are not homogeneous. The simplest case of data heterogeneity is restricted to alternative mark-up syntax because of the use of different elements for the same information. To solve such discrepancies, a recommendation/best practice for the TEI P5 manuscript module (i.e. TEI P5 schema for ENRICH) has already been created and will be tested in the course of the ENRICH project. On the other hand, there are more difficult tasks to be solved. Different formats are usually based on different data models, which are often based on different conceptual frameworks, because they are intended for different goals and purposes. Of course, it is quite possible to convert various formats into the uniform internal format of Manuscriptorium but sometimes with a loss of information and sometimes without any loss of information, however with little correlation to the other data or metadata in terms of their content. Naturally, the loss of information concerns not the loss of the actual text but only the loss of the deep text structure of the descriptive record. On the other hand, the conversion of text into such indistinct fields/elements as <p> or <note> is not very helpful, because it entails the loss of definite meaning, which is so necessary for search. Thus, searching according to the selected fields/elements can sometimes make very little sense, and this challenge is not easy to answer. A methodology other than the mere automatic conversion of data formats into a uniform internal format would have to be used for a successful solution. Such a methodology must first be discovered and then tested before it can be applied.

Moreover, there are further challenges. In the past stages of language evolution, language systems not only in Latin but especially in vernacular languages lacked any conception of strict prescriptions, i.e. languages in the past used neither a prescriptive grammar nor a normative orthography. However, the usual description of manuscripts is based on the use of *rubrics*, *incipits*, *explicits*, *colophons*, etc.—in other words, it is based on quotations from historical texts using their original wording. Consequently, no standard text search (based on the idea of strict prescriptions) can be implemented effectively and successfully. Thus, applying such a technique and methodology as tolerance (variations of one character written differently or erroneously) is not sufficient for a digital library dealing with manuscripts. A more sophisticated technique and methodology like e.g. graphical variants (for texts without normative orthography) must be used for the search in a digital library dealing with manuscripts to be effective and successful. A list of graphical variants, i.e. equivalents, for appropriate languages must be created and implemented into the search system of a digital manuscript library. At present, graphical variants for Latin, Czech and German have been implemented within the Manuscriptorium search system. As for graphic variants, they are represented for each language in a relatively independent module; future extensions are possible.

Another challenge is multilingualism. It can be understood in three ways. Firstly, it is a simple localisation of the user interface in various languages. It will be created during 2009 in the context of the TEI P5 transformation of the Manuscriptorium system (the user interface is currently localised only in Czech and English). Although such a kind of multilingualism is comfortable for users, it is not advisable for scholarly or any other refined use, because it has no relation to deeper search possibilities. Secondly, multilingualism is a translation of both individual descriptive records and a list of search results from one specific language into any other. It can be of great importance if an individual user does not know the language of the manuscript record; however, it is not advisable for scholarly or any other refined use because that use is based on a knowledge of the document language which is mirrored in the descriptive record. On the other hand, the ability of the Manuscriptorium information system to translate descriptive records as well as search results is a necessary precondition for a more sophisticated implementation of multilingualism, if such an ability will be developed within the ENRICH project. Thirdly, such a multilingualism that is appropriate for scholarly and any other refined use is the multilingual search, i.e. it is enough to submit a query in one particular language and the search works also for any other language. It is an extremely complex issue for every digital library and especially for a digital manuscript library because of the use of various historical phases of individual vernacular languages. Although it cannot be implemented into the Manuscriptorium system in the foreseeable future, it is a long-term goal for the Manuscriptorium team.

Last but not least, ontologies present a challenge. No matter what most librarians and information scientists think, the methodology of the ontologies is not a simple information retrieval language. It is rather a semantic networking of texts respectively fulltexts (as which the manuscript descriptive record can also be understood). The European VICODI project (VIsual COntextualization of DIgital content) was led by the Latvian company RIDemo in 2002–2004 and, with the participation of the National Library of the Czech Republic, it elaborated an ontological system for history using concept flavours (abstract notion, event, individual, location, object, organisation), social group roles (i.e. abstract ideas like person, role and symbol) and time intervals on the one hand and instances (a specific specimen of the abstract concept) on the other. Thus, firstly a flavour is connected with the appropriate role and time interval, i.e. a concrete concept-instance is created, and secondly several flavours are connected into the appropriate semantic net, i.e. a concrete context is created. For a digital manuscript library's practical use, the concept instances are mostly personal and/or place names. On the other hand, the implementation of the VICODI ontologies into the Manuscriptorium search system is very difficult, forcing the work on implementation to be done step by step. Firstly, we wish to use ontologies for the translation of descriptive records and search results (words that should not be translated). This step should be implemented during the progress of the ENRICH project. Secondly, an analysis of the implementation of ontologies must be elaborated, which will come after ENRICH if sufficient funding is raised. Thirdly, the implementation of VICODI ontologies into the Manuscriptorium search system depends on the results of the previous analysis.

2.3 Virtual Collections and Virtual Documents

Not only do information and communication technologies enable the overlapping of space and time by sharing different collections held at different places worldwide in a uniform user interface, but they also make it possible to approach the manuscripts and manuscript collections virtually. They allow for the representation of what has no actual existence but what existed either in the past or exists only at an abstract level. Thus, the reconstruction of dispersed historical libraries, the reconstruction of an activity of an individual scriptorium, the reconstruction of a work of an individual scribe, a palaeographic anthology etc. can be accomplished. It is fundamental that all these and other possible reconstructions be accomplished not by accumulating descriptive records, which would be no difference in comparison with the printed environment, but that it be achieved by collecting digital documents, which is very important as against the method not using information and communication technologies. Some virtual collections, e.g. collections of manuscripts, incunabula, early printed books etc., meet typical end-user requirements and as such should be created beforehand, *a priori.* On the other hand, some virtual collections should be created not only *a posteriori* but also for specific purposes and individual tasks. Furthermore, also virtual documents should be created *a posteriori* for specific purposes and individual tasks.

Consequently, personalised research tools for creating virtual documents and virtual collections will be developed in the ENRICH project. A virtual collection can be either static or dynamic. A static virtual collection is a simple selection of documents from those in the Manuscriptorium database at the moment, which requires manual work. A dynamic virtual collection is a selection of documents from the documents in the Manuscriptorium database at that moment or that will be there in future. Such a collection is not based on manual work but on a personalised permanent query that is activated after every update of the Manuscriptorium database, so that the dynamic virtual collection is continuously growing. Especially this kind of virtual collection could be applied well for scholarly and any other refined use, because it enables permanent heuristics. It is very important that static and dynamic virtual collections can be both used individually by their creators and shared with other end users in accord with their creators' consent. On the other hand, a virtual document is a selection of parts of the documents from all the documents currently in the Manuscriptorium database, i.e. manual work is needed for the creation of the virtual document. Practically, it means that a virtual document is a collection of pages or leaves that come from various documents, i.e. manuscripts and/or books. It is thus a specific heuristic tool, whose desirability and efficiency should be proved in the near future.

3 Conclusion

The Manuscriptorium Digital Library uses the ENRICH project to accelerate its development in becoming an important tool for codicological and palaeographic research. The integration of various resources from many European countries as well as the development of tools that would enable the processing of distributed compound digital documents and/or scholarly work with manuscripts are the main advantage of the Manuscriptorium when compared with other similar resources.

Bibliography

Cummings, James and Lou Burnard. *Article on TEI P5 development and use in frame of ENRICH project.* <http://enrich.manuscriptorium.com/index.php?q=node/50>.

ENRICH: *European Networking Resources and Information concerning Cultural Heritage.* <http://enrich.manuscriptorium.com>.

ENRICH: *Definition of basic conditions for sharing of large data sets in the frame of Manuscriptorium.* <http://enrich.manuscriptorium.com/files/ENRICH_WP5_D_5_1_final.pdf>.

ENRICH: *Description of the standards used by the partners, definition of collaboration principles, data and metadata standards.* <http://enrich.manuscriptorium.com/files/Enrich_D2%202_final.pdf>.

ENRICH: *Description of Work.* <http://enrich.manuscriptorium.com/index.php?q=system/files/ENRICH_Grant_Agreement_Annex_I_Description_of_Work.pdf>.

ENRICH: *Report on Development validation of Migration Tools.* <http://enrich.manuscriptorium.com/files/ENRICH_WP3_D3_3_Migration_Tools_01.pdf>.

ENRICH: *Report on pilot full integration and publication of selected partner's metadata in Manuscriptorium.* <http://enrich.manuscriptorium.com/files/ENRICH_WP5_D_5_2_final.pdf>.

ENRICH schema. <http://tei.oucs.ox.ac.uk/ENRICH/Deliverables/referenceManual_en.html>.

ENRICH: *Survey results and their interpretation.* <http://enrich.manuscriptorium.com/files/ENRICH_WP2_SurveyResults_01_01_0.pdf>.

Manuscriptorium: Manuscriptorium Digital Library. <http://www.manuscriptorium.com> or <http://www.manuscriptorium.eu>.

MASTER: *Manuscript Access through Standard for Electronic Records.* <http://www.tei-c.org.uk/Master/Reference/oldindex.html>.

MASTER+. <http://digit.nkp.cz/MMSB/1.1/obr1.bmp> <http://digit.nkp.cz/MMSB/1.1/msankaipXSDdocumentation.html> <http://digit.nkp.cz/MMSB/1.1/msnkaip.xsd>.

TEI P5: *Guidelines for Electronic Text Encoding and Interchange. Chapter 10: Manuscript Description.* Ed. Lou Burnard and Syd Bauman. Oxford, Providence, Charlottesville, Nancy. 2008. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html>.

TEI-aware Software. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/USE.html#CF>.

Toffler, Alvin. *The Third Wave.* Toronto: Bantam Books, 1981.

VICODI: VIsual COntextualization of DIgital content. <http://www.vicodi.org/about.htm>.

Representation and Encoding of Heterogeneous Data in a Web Based Research Environment for Manuscript and Textual Studies

Daniel Deckers, Lutz Koch, Cristina Vertan

Abstract

This paper describes the general architecture of a digital research environment for manuscript and textual studies (particularly those pertaining to ancient Greek and Byzantine texts), and it discusses some questions of data representation and encoding in the framework of such an online research platform. The platform is being developed by the project *Teuchos. Zentrum für Handschriften- und Textforschung*, established in 2007 by the *Institut für Griechische und Lateinische Philologie* (Universität Hamburg) in cooperation with the *Aristoteles-Archiv* (Freie Universität Berlin). Teuchos is a long-term infrastructural project of the *Universität Hamburg*. It is currently in its three-year initial phase which is being co-funded by the German Research Foundation (DFG) through the "Thematic Information Networks" scheme within the "Scientific Library Services and Information Systems" programme. We introduce the main object types to be handled by our system and describe the overall functionality of the online platform. The paper focuses on the representations of two main object types: manuscripts as textual witnesses and watermarks, with an emphasis on the former. Since the adequate encoding of different layers of structure of a transmitted text is particularly relevant to optimising users' choices of navigating both digital images of the containing manuscripts and trancriptions of the text contained, this topic is discussed in some detail. We introduce the formal data model and the corresponding encoding for the object types discussed. The project encodes textual data in XML, aiming for TEI conformance where possible. Since no accepted XML model exists for the encoding of metadata within a watermark collection, we briefly explain how we chose to model the objects to accomodate the collections the project is making accessible.

Zusammenfassung

Der folgende Aufsatz beschreibt die Gesamtarchitektur einer digitalen Arbeitsumgebung für Handschriften- und Textforschung (insbesondere im Bereich altgriechischer und byzantinischer Texte) und Lösungsansätze zu einigen Problemen der Datenrepräsentation und Kodierung im Rahmen einer solchen Online-Plattform. Die Plattform

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 79–95.

wird durch das 2007 am *Institut für Griechische und Lateinische Philologie* (Universität Hamburg) gegründete Projekt *Teuchos. Zentrum für Handschriften- und Textforschung* in Kooperation mit dem *Aristoteles-Archiv* (Freie Universität Berlin) entwickelt. Teuchos ist als langfristige Infrastruktureinrichtung der *Universität Hamburg* angelegt und befindet sich derzeit in seiner dreijährigen Startphase, die von der Deutschen Forschungsgemeinschaft (DFG) im Programm »Themenorientierte Informationsnetze« des Förderinstruments »Wissenschaftliche Literaturversorgungs- und Informationssysteme (LIS)« durch eine Anschubfinanzierung mitgetragen wird. Wir stellen zunächst die wichtigsten Arten von Objekten vor, die das System verwendet, um dann die übergreifende Funktionalität der Plattform zu beschreiben. Der Schwerpunkt liegt hier auf der Darstellung zweier zentraler Objektarten: Handschriften in ihrer Funktion als Textzeugen sowie Wasserzeichen, wobei die Handschriften ausführlicher behandelt werden. Insbesondere wird auf ein geeignetes Kodierungsmodell für verschiedenartige Strukturierungsebenen handschriftlich überlieferter Texte eingegangen, da ein solches von zentraler Bedeutung ist, um den Nutzern möglichst vielseitige Möglichkeiten zu bieten, einerseits durch die Digitalisate der texttragenden Handschriften, andererseits auch durch die Transkriptionen der enthaltenen Texte zu navigieren. Formale Datenmodelle und die zugehörige Kodierung für die behandelten Objektarten werden kurz dargestellt. Innerhalb des Teuchos-Projekts werden Textdaten in XML kodiert, soweit möglich TEI-konform; da kein etabliertes XML-Modell für die Kodierung von Metadaten innerhalb einer Wasserzeichensammlung existiert, wird zudem die Objektmodellierung für die durch das Projekt online bereitgestellten Sammlungen skizzenhaft erläutert.

1 Introduction

This paper briefly describes the general architecture of a digital research environment for manuscript and textual studies, as well as discussing some questions of data representation and encoding. The project *Teuchos. Zentrum für Handschriften- und Textforschung* was initiated in 2007 by the *Institut für Griechische und Lateinische Philologie (Universität Hamburg)* in cooperation with the *Aristoteles-Archiv (Freie Universität Berlin).* Teuchos is a long-term infrastructural project of the *Universität Hamburg* that is currently in its three-year initial phase which is being co-funded by the German Research Foundation (DFG) through the “Thematic Information Networks” scheme within the “Scientific Library Services and Information Systems” programme.

In its final form Teuchos is to provide a web based research environment suited for manuscript and textual studies, offering tools for capturing, exchange and collaborative editing of primary philogical data. The data shall be made accessible to the scholarly community as primary or raw data in order to be reusable as source material for various individual or collaborative research projects. This objective entails an open access policy using creative commons licenses regarding the content generated and published by means of the platform (esp. digital images of manuscripts may have to be handled restrictively dependant upon the holding institutions' policies). The software developed in the course of the project will be made available under free open source licensing as a contribution to the evolving diversity of digital humanities tools and applications.

Distinctive features of the Teuchos platform are the integration of heterogeneous research data (cf. 2) and the participation of different user groups in the generation and enhancement of the content. The system as a whole is geared to the needs and preferences of specialised research (rather than to the presentation of library treasures to a wider public). The following use cases are forseen:

- Provision of data facilitating the use of digitised manuscripts (created and shared by different user groups), ranging from structural information regarding the intellectual content of the manuscript to transcriptions containing indications of variant readings.
- Provision of digitised manuscripts accompanied by (partial) transcriptions both as a basis for further editorial work and to make core information on the content and the manuscript tradition available (and citable) to the scholarly community at the same early stage. While in certain cases this first step may be the only one that will be taken, with a view to the other cases it may also be considered a methodological improvement for textual studies in general to render the separate stages of the editorial process verifiable.
- Collaboration of networked researchers independent of time and space as a prerequisite for the analysis and utilisation of special materials, e.g. domain-specific texts and inaccessible or damaged sources such as palimpsests.
- An evolving collection of manuscript descriptions gives access to detailed information on codicology, manuscript history and textual transmission. This material derives from autoptical library studies and is thus often inherently sporadic and disjointed; on the other hand the collection is independent of library catalogisation projects and open to the collaboration of researchers worldwide who contribute according to their respective field of expertise and/or their serendipitous findings.
- A flexible model allows for the integration of manuscript descriptions of varying depth. A substantial amount of material taken from both published and unpublished materials of the *Aristoteles Graecus* offers a model for comprehensive and highly structured descriptions. The complex relationships between manuscripts belonging to a coherent textual tradition, e.g. common sources, scribes, owners, annotators etc. offer multiple possibilities of inter-linkage, which may be used for the individual exploration and eventually for automated analysis of such a corpus.

- Value is continuously added by way of cross-linking and data exchange with existing online resources relevant to manuscript and textual studies, especially library catalogues and manuscript bibliographies, but also archives of texts and digital libraries.

A prerequisite for the intended uses of the Teuchos platform are open data formats conformant to established community standards, the most important for several areas being TEI P5. Standards and data models to be used were also chosen with a view to the later addition of mechanisms to facilitate interoperability on the level of both basic metadata (e.g. through support of OAI-PMH) and of complex data sets (e.g. automatic conversion of the larger part of manuscript description data to the Manuscriptum XML format).

Following an overview of the system in general, we will proceed to discuss some specific encoding issues arising from the combination of materials we provide, and which we believe to be somewhat particular.

2 Teuchos Functionality and its Use for Codicology and Palaeography

The system we describe is meant to offer scholars and editors in the field of Greek codicology and palaeography a powerful tool for creating, publishing and augmenting relevant research materials of the kinds described below, as well as offering extensive search options for the larger scholarly community interested in consulting such materials. An overview of the basic system architecture is given in Figure 1.

All our objects are stored in a Fedora repository. This technology was chosen both for reasons such as standards conformance, robustness, openness etc., and because it has a wide user base in the eSciences and eHumanities.

The user interacts with this repository via a web application that manages the editing, searching, and uploading processes. We have three general categories of users:

- System administrators with full access to all parts of the system. They exercise overall control of the content made available by the other users.
- Registered users have access to released material on the system, and in addition they can contribute by uploading materials, writing forum entries and accessing and editing both released and preliminary materials within the scope of their respective user group. We envisage a hierarchy of such users having different access types to various parts of the system.
- Public users, who can only view released materials that carry no additional restrictions.

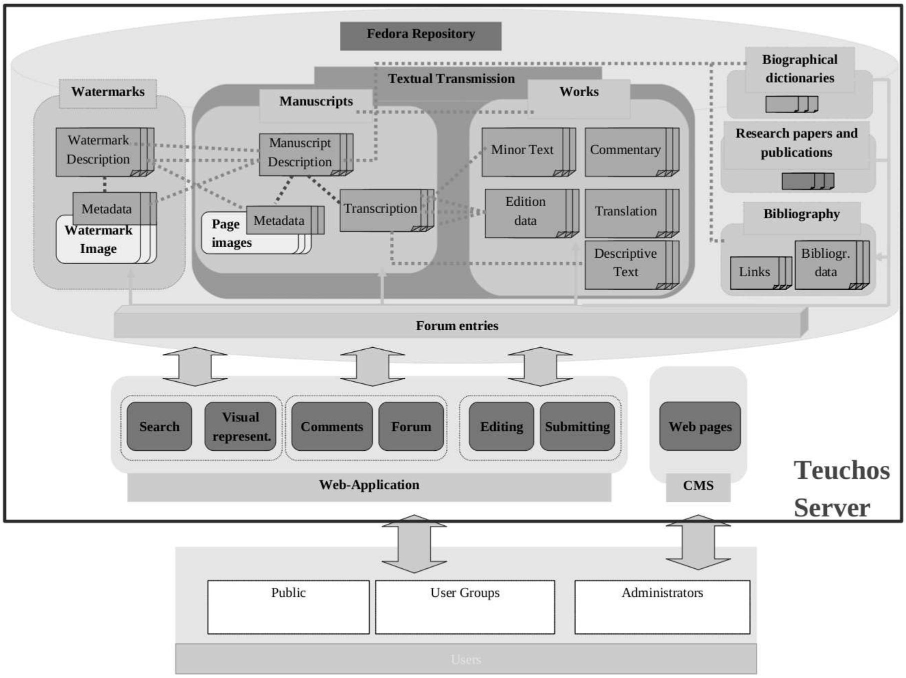


Figure 1. Teuchos Architecture.

There are several groups of digital objects to be stored in the Fedora repository:

- We store tracings of *watermarks* from dated paper manuscripts as digital images on the one hand, and descriptive data on these watermarks and their motif groups in an XML format on the other. Images are associated with Dublin Core (DC) and additional similarly structured metadata and linked to the descriptive data. A more detailed explanation is provided in 3.3.
- The *textual transmission* group is divided into two subgroups that are themselves subdivided: material related to individual *manuscripts* and material related to a particular *work*, e.g. a particular source text by a particular author.
- The *manuscript* group encompasses digital *page images* of manuscripts (or parts of manuscripts) that are aggregated on a per manuscript basis (cf. 3.1), scholarly *manuscript descriptions* that may reference page images if available for the manuscript described, and *transcription* data, which may range from a first set of basic structural data (cf. 3.2) to full transcriptions, and usually links to pages of ex-

actly one manuscript (exceptions are e.g. texts spanning more than one manuscript volume and re- or misbound manuscripts).

- The group of *works* encompasses a wide range of materials refering to a source text with its entire set of manuscripts rather than to one particular witness, and ranges from *full critical editions* (with several intermediate stages) and *translations* to various kinds of *commentaries* (and other explanatory or *descriptive* materials).
- The *biographical dictionary* group hosts prosopographic information on historical persons relevant to our other research materials.
- A special group is dedicated to *research papers* that may reference material from the other groups, without themselves falling into any of the other categories.
- Finally, *bibliographic data* (also including online resources) pertinent to specific research areas is collected and made available as a separate group of objects.

The Fedora repository offers a flexible mechanism to store heterogeneous information for one and the same object. A digital Fedora object consists of one unique identifier and several data streams. This facilitates the handling of these objects as basic relationships between objects and basic DC data for each object stored in separate streams from the main XML or image data as described above. More detailed information on individual authorship and responsibility for any particular part of one textual object produced cooperatively by a group of scholars, which is annotated upon data entry, can thus also be stored separately in an additional data stream. Should later extensions to the project result in additional kinds of annotation (e.g. linguistic or semantic), these could also easily be added as stand-off markup in separate streams without repercussions on the main stream's data format.

3 Data Representation and Encoding

Due to the heterogeneous nature of the data in the Teuchos platform, only part of it can be encoded in XML following the TEI P5 guidelines, most notably manuscript descriptions and transcriptions. The latter will be discussed at length later in this chapter. Especially where we handle digital images, further structures besides the means of reference foreseen in TEI are required, and collections such as e.g. the watermark albums are outside of its scope of representation. Since not all materials will be available for all manuscripts, the platform is implemented as a robust system able to accomodate incomplete sets of data.

3.1 Manuscripts: Digital Images

Not all descriptive materials in Teuchos are accompanied by digital images of manuscripts. In the cases where we can provide such images, each of these is ac-

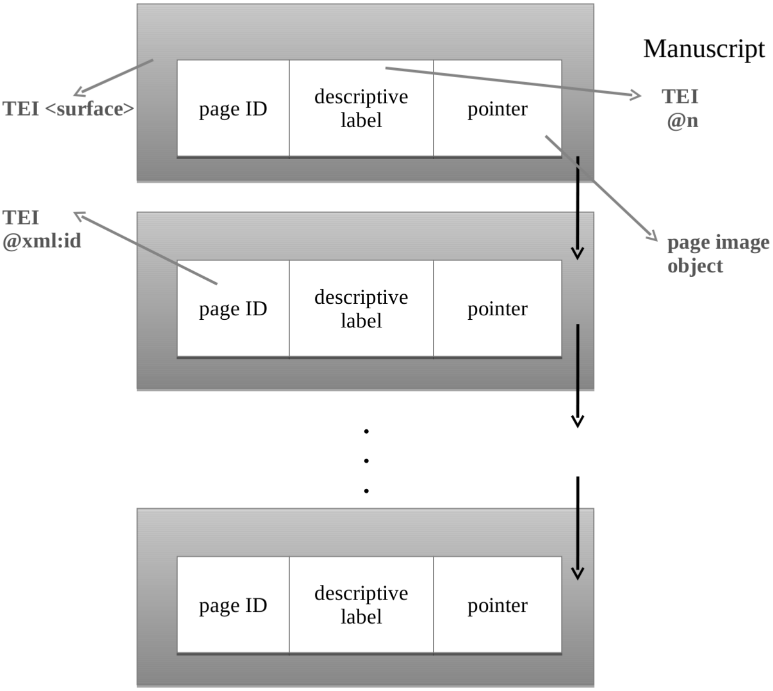


Figure 2. Linking page information within a manuscript object.

companied by a set of descriptive metadata as well as authority data. This per page image metadata will be stored in a Fedora object together with a thumbnail of the image for faster access e.g. for search processes and preview purposes.

For each manuscript of which we provide digital images, a reference document making use of the TEI <facsimile> element is created. In our case, this consists solely of a list of <surface>s, providing a unique identifier in the form of an xml:id attribute, and a label discussed below by using the n attribute. For each surface, a pointer to the Fedora object just mentioned containing a thumbnail and reference to the full image data is added.

We try to provide mostly unambiguous human readable labels for all pages (usually the main foliation or pagination and the customary short designations also used in manuscript descriptions for flyleaves etc.). Where deemed necessary (e.g. for images of the binding), a more descriptive label to be used in the display of images might be additionally supplied using the <desc> element.

To ascertain that the page structure of the manuscript at the time it was digitised is represented, <surface>s are listed according to their physical order in the manuscript, and in the case of pages not digitised, an empty surface element will be included (since we do not carry out all digitisations ourselves, there have been cases where e.g. flyleaves have been omitted).

While the use of this minimal <facsimile> document is quite limited, it separates information on image files; their location and availability from the structural data contained within the transcriptions discussed in the next chapter, and at the same time holds the basic information required for a simple browsable presentation providing minimal navigational aids. This ensures both that the image material can be made accessible independently of the (partial) completion of a transcription (and thus prior to any in-depth analysis of the textual content), and that no images of matter that is deemed to be outside the scope of the transcription are excluded through mere omission of reference.

If images of manuscripts not available through Teuchos are made individually accessible by a third party, reference to such images could be made through this same abstraction mechanism.

Manuscripts: Physical and Textual Structure

To facilitate users' access to digital images of a manuscript, we also provide at least a minimal (i.e. possibly empty except for pagebreaks) transcription of the manuscript, containing structural information that can be used to offer alternate representations and improved navigation for browsing, and to give a clearer indication of the part of the text to which an image viewed pertains. Such data is encoded within a TEI <text> element. Similar information will of course also be included in full transcriptions even of manuscripts for which we cannot provide images, but we will limit ourselves to discussing merely the structural elements in this paper.

We should like to note that the structural data we discuss in this section does not necessarily presuppose the existence or public availability of images of the manuscript or manuscripts referenced. While our basic structural data can with relative ease be extracted to a standard format such as METS for export, this is only meaningful where digital images can be referred to, and information on the precise placement of structural elements within a (partial) transcription is still lost this way. In this regard, our approach is orthogonal to models that treat the page image as the main point of reference for the structural metadata.

Page, Folio or Quire Structure

Reference to the digital images of individual manuscript pages is made through the inclusion of <pb> elements using the corresp attribute to point to the unique identifiers established for the representation of each page (cf. all three examples below). Foliation or pagination information, however, is encoded separately using the <fw> element. This permits recording whether numbers provided by the transcriber are actually present on the page or not, and also recording more than one such reference system (besides misfoliation and the like, the presence of a second, alternate foliation is not uncommon) through consecutively including more than one <fw> element (cf. ex. 1 and 3). There is the slight problem of clearly disambiguating multiple foliation systems. While the difference is usually evident from the placement attribute (though one may have to introduce very precise values for this attribute e.g. to clearly separate two foliations indicated one above the other), the use of this positional data for classification is less than ideal. Since alternate numbering schemes will usually be supplemented with some typographic symbol by the scholar working on the description or transcription of a manuscript to make them easily distinguishable to users of the materials, our solution is to always supply such 'normalised' forms of any numbering scheme in the n attribute. For verso pages that usually show no numbering, an empy <fw> element is used to supply the number.

Since the <facsimile> element's structure is thus kept independent from that of the transcription, the transcription can without further ado ignore the manuscript's current page order where it does not correspond to the original arrangement (e.g. through misbinding or in the case of the underlying text in palimpsests). Additional information on a manuscript's structure can also be included. Most notably, further <pb> elements may be included for (physically, not digitally) missing pages with or without loss of text (information that is not available within the limits we imposed on the <facsimile> structure above). Where a manuscript exhibits quire signatures, we also include these using <milestone> elements (with a value of "quire" for the unit attribute).

Content-related Structure

The most important means of providing structural information within the scope of a transcription are (1) relating the text to the structure of an existing edition of the work in question, and of course (2) encoding any structure evident in the witness being transcribed itself. Whereas in the former case, these structures should not be encoded hierarchically for evident reason (cf. below), a structure derived from the witness itself would prima facie be deemed safe for hierarchical encoding. However, this no longer necessarily holds true once transcriptions from more than one witness of the work in question are to be joined, e.g. where chapter divisions are inconsistent among manuscripts as one of the most obvious cases. To be able to retain per-witness structural information in a joined document, we therefore propose to encode all structural information using empty elements, i.e. <milestone>s. When such a joined document is edited further to become a new edition of a work in its own right, the editor(s) may (and in most cases probably will) of course decide to create a hierarchical structure taking into account the structure of the various witnesses, but this should be a later step. To avoid confusion, we should state that we do not intend to provide dynamically generated editions. While the semi-automatic joining of transcriptions is a first step towards creating a digital critical edition, the further steps require substantial scholarly intervention.

Content-related Structure Derived from Editions

It is established practice to use structural information derived from an existing edition for new work on a text. While such structures may in whole or in part be of a hierarchical kind (e.g. book, poem and line number for the works of a poet), this is not always the case, be it that the final unit is the line number in a prose text (which might switch in the middle of a word where hyphenation is used), or that the entire reference is based on pages and lines of a previous edition (e.g. for most of Aristotle's work reference is to the pages, or rather columns, in Bekker's edition, followed by a line number). Most of these reference systems use two to three hierarchical levels. We propose the use of <milestone>s using a special value of "external" for the unit attribute to prevent confusion with any intrinsic references, and a special value of "canonical" for the type attribute (cf. ex. 2). The edition and numbering scheme being referred to will be indicated using the ed attribute (thus including references to more than one edition or to multiple numbering schemes derived from one edition is possible), and the hierarchical level using the subtype attribute, for which we propose using a simple range of values "level1", "level2" etc. to facilitate processing. The actual reference is supplied through the n attribute (while references are usually some kind of numerical value, extrinsic short titles or headers might in fact be encoded this way, too). Obviously information on what each value used for the ed attribute refers to and what kind of reference (book, page, stanza, line etc.) the various subtype attribute values represent for this edition need to be supplied in an <encodingDesc> so the information can be displayed in a meaningful way. The granularity of the information encoded is up to the transcriber, and might range from a mere indication of chapters, poems or similar through indication at the precision of a line number at the beginning of each page to a per-line matching between text transcribed and canonical edition. As a special case example, we have provided references to line numbers for an (otherwise unstructured) new version of a text for approximately every 5th to 10th line where this version actually (partially) corresponds to the edition of the previously known versions (cf. ex. 3).

Content-related Structure Derived from the Witness

The same system may be used for structure derived directly from the witness (e.g. book, poem or chapter numbers, cf. ex. 1 and 2). Due to the problems of joining transcriptions discussed above, the values suggested for the unit attribute in the TEI guidelines should again be shunned, and we propose use of the special value "internal" instead, of values "present" (for numbers actually present in the text) and "implied" (for those numbers supplemented from the evident sequence even though they are not or no longer evident in the witness) for the type attribute, again indicating the hierarchical level (where applicable) through the subtype attribute, and of course adding the appropriate reference identifiers to the <encodingDesc>. (And further noting that in spite of its name, the ed attribute can be used to indicate a manuscript siglum instead of an edition in this case. Should a manuscript contain more than one such numbering scheme, a concurrent scheme would have to be indicated through the use of a modified siglum). This intentionally does not take into account where and how the given numbers are actually indicated in the manuscript, as we expect to indicate these numbers in a normalised form. (They might be written textually rather than as numbers, or in Greek or Roman numerals, at the front or at the end of a title, in the margin or in some other location, which would all yield different representations in the transcription.) It is important to note that our approach separates the numbers from textual titles of headings (on the latter cf. below), thus allowing us to gracefully handle the case where identical titles received different numberings in different manuscripts.

Headings and Other Non-numerical Indicators of Structure

Where a manuscript contains chapter (or book, paragraph etc.) headings, we encode these using the <head> element (cf. ex. 1), using an appropriate type attribute to identify the kind of scheme (and to distinguish e.g. a spelled out numbering system from the written titles, if so desired), employing the same subtype and n attributes as with the corresponding numbering scheme (one might argue that this information is superfluous, as we always include a milestone for consistency anyways). Generally, a suitable numbering scheme as described above will be added in case no numbering is employed in the manuscript. In cases of a manuscript mostly following the textual structure established in an edition of the same text from other sources, but omitting the headings, these can usually be included as <supplied> within a <head> element. Should the manuscript being transcribed contain the headings in a nonstandard form (relative to the tradition of this text, or possibly to an established edition), it may be desirable to offer the standardised form, in which case the <head> would contain an <app> element with the standardised form indicated as the <lem>. (We should clarify that we use the parallel segmentation approach for all our transcribed or editoral texts, thus these elements will always be inlined, facilitating extraction of this information.) If the titles are regarded as sufficiently universal, the approach utilising <lem> might conceivably be used even for cases of multi-versioned transmission where no base text is chosen outside the scope of the headings. An alternate approach for this latter case would be to encode such titles as an external reference system (cf. above). We maintain that this latter method is always preferable when supplying typified titles (cf. ex. 2), which might be in a (modern) language other than that of the text (e.g. where the actual titles are incomplete or may be deemed misleading).

To address a case that might be seen as somewhat controversial, in one manuscript with otherwise very little structure, we have one or two phrases written in red ink on almost every page, but few of them would be considered headings, the others being descriptive image labels, minor notes on the content and other materials. Given uncertainties in easily achieving a reliable classification in this particular case, we have as a first step added transcriptions of these phrases tagged merely with <hi>, which can thus be extracted and used as an additional 'navigational' aid for a reader trying to locate some particular part of the text in this manuscript (cf. ex. 3). Obviously proper classification of each of these items according to their content would be preferable, and arguably the approach used here will yield unsatisfactory results for the majority of similar cases. While we have included this possibility to allow users to quickly add a first transcription of highlighted parts of a text in order to make these navigable, we certainly recommend only using it where a proper classification of this material cannot be swiftly achieved, and we very much advise against using this for textual elements that would be considered actual section headings of any kind.

Examples

```
<pb corresp=”#berolham512—165r”/>
<fw n=”186*” placement=”top right upper”>186</fw>
<fw n=”165” placement=”top right lower”>165</fw>
<milestone n=”10” unit=”internal” type=”implied” subtype=”level1”/>
<milestone n=”1” unit=”internal” type=”implied” subtype=”level2”/>
<milestone n=”3” unit=”internal” type=”present” subtype=”level3”/>
<head n=”3” type=”numbering” subtype=”level3”>κεφ. γ′</head>
<head n=”3” type=”title” subtype=”level3”>Περὶ τῆς ὑλικῆς αἰτίας καὶ τῶν περὶ αὐτῆς δοξῶν τῶν
παλαιῶν</head>
```

Example 1. Otherwise empty transcription of one page containing numbering from two foliations, position within a three-level structure (book, title, chapter — not evident) and the current chapter's numbering and title as written on the page.

```
<pb corresp=”#berolphill1538—326r”/>
<milestone n=”Krampfanfall und heilige Krankheit” unit=”external”
type=”titletypefied”/>
<milestone n=”108” ed=”edoh–c” unit=”external” type=”canonical”/>
<milestone n=”109” unit=”internal” type=”present”/>
<fw n=”326” placement=”top right”>326</fw>
<milestone n=”369” ed=”edoh–p” unit=”external” type=”canonical”/>
<pb corresp=”#berolphill1538—326v”/>
<fw n=”326v”/>
<pb corresp=”#berolphill1538—191r”/>
<milestone n=”Hufleiden” unit=”external” type=”titletypefied”/>
<milestone n=”109” ed=”edoh–c” unit=”external” type=”canonical”/>
<milestone n=”110” unit=”internal” type=”present”/>
<fw n=”191” placement=”top right”>191</fw>
<milestone n=”370” ed=”edoh–p” unit=”external” type=”canonical”/>
```

Example 2. Multipage excerpt (note transposition of folios) from an otherwise empty transcription providing typefied modern chapter titles, divergent chapter count from edition and witness transcribed, as well as a reference to the edition's page count (not congruent with the witness' page structure).

```
<pb corresp=”#lipsgab19—46r”/>
<fw n=”σϛ” placement=”top right upper”>σϛ</fw>
<fw n=”46” placement=”top right lower”>46</fw>
<lb n=”1”/><hi rend=”red”>πῶς ἕλληνες ἀπέκλεισαν τὴν τρώϊαν ἐκείνην·</hi>
<lb n=”2”/><milestone n=”10516” ed=”edpj” unit=”external” type=”canonical”/>
<lb n=”3”/>
<lb n=”4”/>
<lb n=”5”/>
<lb n=”6”/>
<lb n=”7”/>
<lb n=”8”/><milestone n=”10528” ed=”edpj” unit=”external” type=”canonical”/>
<lb n=”9”/>
<lb n=”10”/><hi rend=”red”>ἀρχὴ τῆς ὑποθέσεως τοῦ ὁρωμένου κόσμου:</hi>
```

Example 3. Beginning of a page of a transcription from a manuscript with two foliations, unclassified highlighted matter and indication of approximate alignment with a reference edition of a divergent version of the text.

3.2 Collections of Watermarks

Watermarks introduced by the paper-makers are evident in virtually all Western paper that has been used as a writing support in medieval manuscripts and documents. Since watermarks changed frequently, and produced paper was used up within a short span of years, these marks can be an important aid in dating manuscripts. To this end, numerous collections of watermarks from dated documents or manuscripts have been created, and many of these have since been digitised and made available online.

A basic discussion of watermark collections can be found in the paper of Wolf in this volume. The Bernstein project, which has undertaken to offer unified access to some of these collections and is working on a refined and more universal motif classification for use in the digital world, provides a by no means exhaustive list of existing collections on its website.

The collections of watermarks the Teuchos project is preparing to bring online contain tracings of marks from Greek manuscripts, and exhibit some specific features. Most importantly, watermark motifs occur in twin pairs due to the production process of the paper. (The marks are created through wire soldered or sewn onto the base wire structure of the molds used, and the established manual paper production process used two such molds in alternation between each sheet of paper.) Since our collections are based not on documents, but on manuscripts that frequently contain a number of consecutive sheets from the same source, they can usually offer both twin instances of a watermark motif. Given that the molds deteriorate over time, in some cases additional instances of a motif may be available where a degradation of some motif part is evident. It may also be possible to identify identical watermarks in other manuscripts.

We chose to encode our collection data in XML, and to model two different object classes in an object oriented approach (cf. Figure 3 and 4). Motif objects hold information pertaining to a watermark motif as a whole, and they form a super class to instance objects that encode information on a single mark as represented by a single tracing. Since countermarks (smaller subsidiary marks placed at a different location of the sheet) are usually traced separately, we also treat these as individual instances of the main motif.

The motif object contains the main identification data on the motif, and lists of the pertaining instances, of similar motifs (in any collection) and of identical motifs (usually in other collections). The instance object contains all the data that will or may be different between two tracings (e.g. originating manuscript with information on dating etc., exact location in the manuscript), as well as links to the motif object, to the actual digital image of the tracing, and to the description of the manuscript the image was taken from. Since most of our collection contains data on the origin of the datings (specifically precise references to manuscript subscriptions) which are not strictly part of the watermark description, we extract this data and automatically create a minimal manuscript description to host that information. A user wishing to verify details on the origin of the dating indicated with a given watermark has easy access through the link to this manuscript description. An example of an instance object based on our XML schema is presented below:

```
<teuwmo:teuwmObj [...]>
<teuwmo:wmIdent wmIsCountermark=”false”>
<teuwmo:wmObjId>TEU_WMDesc_Aiglem2–21.xml</teuwmo:wmObjId>
<teuwmm:wmIdentification>
<teuwmm:wmIdnr>21</teuwmm:wmIdnr>
<teuwmm:wmCollection>Harlfinger</teuwmm:wmCollection>
<teuwmm:wmName>
<wmNameLanguage wmLang=”fr”>Aigle</wmNameLanguage>
<wmNameLanguage wmLang=”de”>Adler</wmNameLanguage>
</teuwmm:wmName>
</teuwmm:wmIdentification>
</teuwmo:wmIdent>
<teuwmo:wmManuscriptData>
<teuwmo:msName>Vatic. 1469</teuwmo:msName>
<teuwmo:msFolio>ff. 1–72</teuwmo:msFolio>
<teuwmo:msDate>1495</teuwmo:msDate>
</teuwmo:wmManuscriptData>
<teuwmo:wmLinks>
<teuwmo:pictureLink>Aigle–21m2</teuwmo:pictureLink>
<teuwmo:msDescLink>Aigle–21.xml</teuwmo:msDescLink>
<teuwmo:motifLink>Aigle.xml</teuwmo:motifLink>
</teuwmo:wmLinks>
</teuwmo:teuwmObj>
```

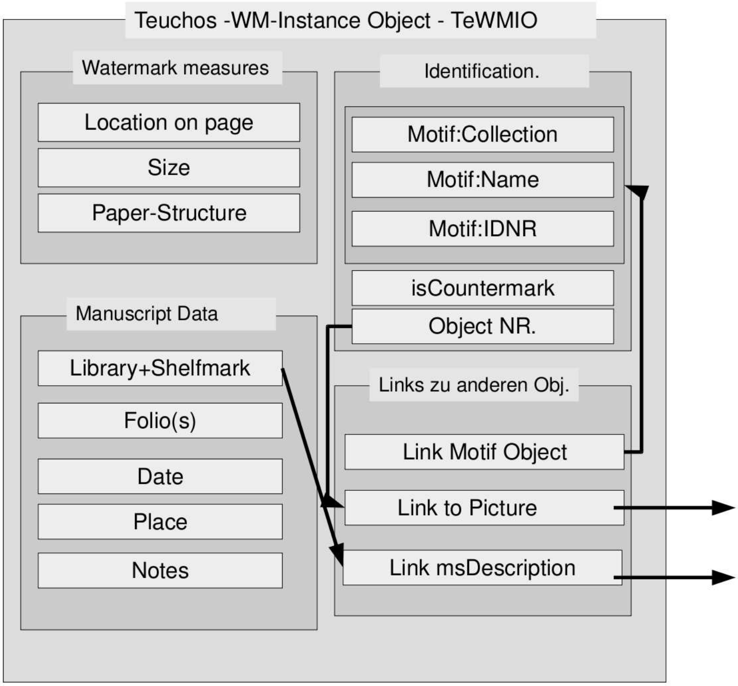


Figure 3. Watermark motif object.

4 Conclusions

In this paper, we introduced the Teuchos center as an evolving research environment for manuscript and textual studies. Due to constraints of space, we gave a general overview ignoring important aspects like user-interfaces, CMS, search functionality etc., and proceeded to discuss specific encoding decisions in selected areas. We chose to focus particularly on structural encoding as a precursor to preparing transcriptions, since we consider this an important step in providing fuller access to digitised manuscripts for textual scholars, and a necessary prerequisite for cumulative and shared scholarly work on the primary text sources in a distributed digital environment. As a consequence of this focus, we did not cover manuscript descriptions, for which we resort to TEI P5 as well. While the conversion of numerous existing in-depth descriptions, and the extensions of the manuscript description provisions of TEI to support all the particular aspects of the *Aristoteles Graecus* cataloguing model, as well as providing for more in-

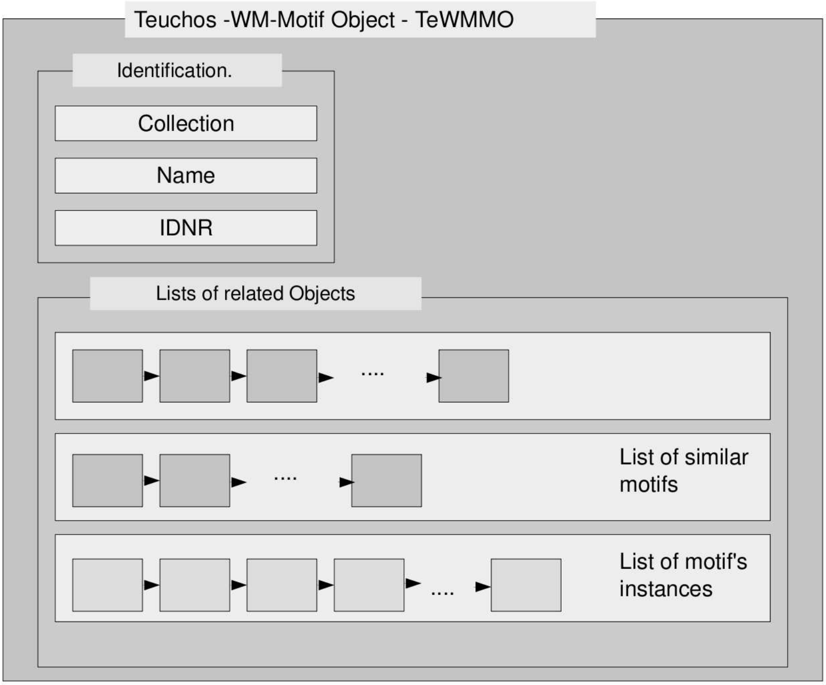


Figure 4. Watermark instance object.

depth analytical markup of the existing and of new descriptions, would be a subject for much discussion, it would have gone far beyond the scope of this paper.

Bibliography

Aristotelis Opera ex rec. Immanuelis Bekkeri. Ed. Academia Regia Borussica. 4 vols. Berlin, Reimer: 1831-1836.

Bernstein project's overview of watermark collections. <http://www.bernstein.oeaw.ac.at/twiki/bin/view/Main/PaperDatabases>.

DFG, Thematic information networks. <http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/projektfoerderung/foerderziele/informationsnetze.html>.

Dublin Core. <http://dublincore.org>.

Fedora. <http://www.fedora.info>.

Moraux, Paul et al. *Aristoteles Graecus. Die Griechischen Manuskripte des Aristoteles.* Berlin: De Gruyter, 1976 ff.

Oder, Eugen and Karl Hoppe, eds. *Corpus hippiatricorum Graecorum*, 2 vols., Leipzig: Teubner, 1924/1927.

Papathomopoulos, Manolis and Elizabeth Mary Jeffreys. *Ο πόλεμος της Τρωάδος.* Athens: Μορφωτικό Ίδρυμα Εθνικής Τραπέζης, 1996.

Stevenson, Allan Henry. "Watermarks are Twins." *Studies in Bibliography* 4 (1951-52): 57-91.

Teuchos. <http://www.teuchos.uni-hamburg.de>.

Examples

Example 1. Berlin Hamilton 512, f. 165r. The part of the manuscript in the example contains a philosophical text book by 13th century scholar Georgios Pachymeres. The two foliations hail from before and after the loss of the first 21 folios of the codex.

Example 2. Berlin Phillips 1538, f. 326r, 326v and 191r. A richly decorated 10th century manuscript containing a collection of texts on horse medicine. The chapters in the example deal with convulsions and epilepsy and with hoof diseases respectivly, reference is made to the edition of Oder and Hoppe.

Example 3. Leipzig Gabelenz 19, f. 46r. A version of the Greek adaptation of the medieval epic poem War of Troy. The manuscript contains an alternate foliation using Greek numerals. The first caption transcribed is descriptive, the second a classification of what is to follow. Reference is made to the verse numbers in the modern edition of the main tradition by Papathomopoulos and Jeffreys.

Aufbau eines Informationssystems für Wasserzeichen in den DFG-Handschriftenzentren[*]

Christina Wolf

Zusammenfassung

Die Württembergische Landesbibliothek, das Landesarchiv Baden-Württemberg, die Bayerische Staatsbibliothek und die Universitätsbibliothek Leipzig planen ein Projekt zur Errichtung eines Informationssystems für Wasserzeichen und deren Beschreibungen. Dieses soll den Aufbau und die Verwaltung digitaler Wasserzeichensammlungen erleichtern und sie über einen zentralen Zugriffspunkt im Internet für die Wissenschaft nutzbar machen. Zunächst ist vorgesehen, ein »Backend« für die dezentrale Dateneingabe zu schaffen, wozu ein bereits existierendes Erfassungsmodul der Österreichischen Akademie der Wissenschaften angepasst und weiterentwickelt werden kann. Die Darstellung der Inhalte im Internet soll über ein Online-Präsentationssystem erfolgen, das aufbauend auf dem bereits bestehenden Frontend von »Piccard-Online« entwickelt und mit zusätzlichen Funktionalitäten versehen werden soll. In das Informationssystem sollen zunächst die Nachweise der »Piccard-Online«-Anwendung und der noch nicht digitalisierten Piccard-Findbücher des Landesarchivs Baden-Württemberg sowie etwa 12.000 neue Wasserzeichen aus den drei oben genannten Handschriftenzentren eingebunden werden. Um eine homogene Eingabe in das Erfassungsmodul zu erreichen und differenziertere Suchfunktionen zu ermöglichen, werden Richtlinien für die Digitalisierung und Beschreibung von Wasserzeichen festgelegt. Weiter ist die Vernetzung mit fachverwandten Systemen wie z.B. »Manuscripta Mediaevalia« und dem »Bernstein«-Portal geplant.

Abstract

The Wuerttembergische Landesbibliothek, the Landesarchiv Baden-Wuerttemberg, the Bayerische Staatsbibliothek and the Universitaetsbibliothek Leipzig are planning a project for building an information system for watermarks and their descriptions. It is supposed to facilitate the creation and management of digital watermark collections and to make them available via a central access point in the internet. The institutions design to create a back-end for the decentralised input of data. For this purpose, an already existing capture tool of the Oesterreichische Akademie der Wissenschaften can be

[*] Der Beitrag ist unter Mitwirkung von Dr. Alessandra Sorbello Staub, Württembergische Landesbibliothek Stuttgart, entstanden.

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 97–107.

adapted and refined. For the presentation of content in the web, an online presentation system will be developed on the basis of the "Piccard-Online" front-end and provided with further functionalities. The first contents for the information system will be taken from the Piccard collection of the Landesarchiv Baden-Wuerttemberg and about 12.000 new watermarks from the manuscript centres mentioned above. It is planned to specify guidelines for the digitization and description of watermarks in order to obtain a unified data input and enable sophisticated searches. Furthermore it is planned to connect the watermarks information system with related systems such as "Manuscripta Mediaevalia" and the "Bernstein" portal.

1 Ausgangslage und erste Planungen

Die Bestimmung von Wasserzeichen gehört zu den Grundlagen der kodikologischen Beschreibung von Papierhandschriften und liefert wichtige Anhaltspunkte für ihre Datierung und ihren Aufbau. Auch bei Fragestellungen in anderen Disziplinen wie der Inkunabelkunde, der Kunstgeschichte oder der Musikgeschichte kann die Untersuchung von Wasserzeichen nützliche Aufschlüsse geben. Neben gedruckten Repertorien – wie vor allem Piccard und Briquet – kann man als Referenz für die Ermittlung von Wasserzeichen auf verschiedene Online-Angebote zurückgreifen. Eine zentrale Rolle für deutsche Handschriften aus dem deutschsprachigen Raum kommt hier der Datenbank »Piccard-Online« zu, einer digitalen Fassung der im Landesarchiv Baden-Württemberg verwahrten Wasserzeichensammlung Piccard, die mit circa 92.000 Einzelbelegen als weltweit größte ihrer Art gilt. Aus knapp 90 Archiven und Bibliotheken hatte Gerhard Piccard (1909-1989) die Belege zusammengetragen und seit 1961 sukzessive etwa die Hälfte davon nach Motiven geordnet in 17 Findbüchern publiziert. Mit Förderung der Deutschen Forschungsgemeinschaft wurde 2003-2006 vom Landesarchiv Baden-Württemberg eine digitale Fassung der Sammlung als Online-Findmittel umgesetzt.

Neben der Piccard-Datenbank sind auch die Datenbanken »Wasserzeichen des Mittelalters« (WZMA), »Watermarks in Incunabula printed in the Low Countries« (WILC) und »International Database of Watermarks and Paper Use for Prints and Drawings c. 1450-1800« (NIKI) zu nennen, die alle über das im Rahmen eines EU-Projekts realisierte Portal »Bernstein – the memory of paper« erreichbar und seit Februar 2009 auch gemeinsam über eine Metasuche recherchierbar sind. Die einzelnen Webangebote sind hinsichtlich der Zusammenstellung der Sammlungen (abgedeckte Regionen, Chronologie), der Datenerfassungsschemata, der Art der Gewinnung und der Bereitstellung der Informationen (Durchzeichnung, Durchreibung, Thermografie, Betaradiografie) unterschiedlich gestaltet.

Mit diesen Angeboten stehen den Nutzern bereits einige Werkzeuge für die reine Recherche in Wasserzeichensammlungen zur Verfügung. Sie werden dabei allerdings mit verschiedenen Hindernissen konfrontiert: Aufgrund der Heterogenität der Systeme und Sammlungen gestaltet sich die Suche in den Beständen mühsam. Sie verläuft nur in einer begrenzten Anzahl von Fällen erfolgreich. Funde in Online-Datenbanken können zudem nicht automatisch und im Sinne einer eindeutigen Referenzierung in elektronische Nachweissysteme für die Katalogisierung von Handschriften und Nachlässen – wie etwa die Verbunddatenbank »Manuscripta Mediaevalia« – übernommen werden. Da es sich bei den digitalen Angeboten im Netz um geschlossene Sammlungen handelt, besteht zudem keine Möglichkeit, Neufunde zu verzeichnen und den vorhandenen Datenbestand anzureichern. An dieser Stelle möchten die Handschriftenzentren[1] ansetzen, die seit langem den Wunsch nach einem offenen, den Bedürfnissen der Handschriften- und Nachlasskatalogisierung sowie allgemein der Papierforschung gewachsenen Nachweissystem hegen. Im vergangenen Jahr hat die Württembergische Landesbibliothek daher im Auftrag der Handschriftenzentren die Planungen für ein Projekt übernommen, in dem ein Informationssystem für die Erschließung der im Rahmen von Handschriftenkatalogisierungsprojekten zusammengetragenen Wasserzeichensammlungen aufgebaut werden soll. Es ist vorgesehen, noch in diesem Jahr mit der Realisierung des Vorhabens im Rahmen eines von der Deutschen Forschungsgemeinschaft geförderten Projekts zu beginnen.

Die Württembergische Landesbibliothek selbst verfügt über Durchzeichnungen von rund 4.000 Wasserzeichen, die Gerhard Piccard für diese Einrichtung angefertigt hat. Da diese Aufzeichnungen jedoch nicht, beziehungsweise nur zu einem geringen Teil, Eingang in seine Kartei gefunden haben, kann man hier weitestgehend von bisher unbekannten Nachweisen ausgehen, die mithilfe des geplanten Informationssystems der Wissenschaft digital zugänglich gemacht werden sollen. Bei Zustandekommen des Projektes wird sich die Württembergische Landesbibliothek über die Aufbereitung und Einbindung dieser Sammlung hinaus an der Festlegung inhaltlicher Anforderungen für das Informationssystem und für die Erschließung der Wasserzeichen beteiligen, sowie an der Koordination und Redaktion der Dateneingabe und der Prüfung von Vernetzungsmöglichkeiten mit bereits bestehenden, thematisch verwandten Systemen mitwirken.

Als Partner für die technische Realisierung und die Festlegung von Erschließungsrichtlinien konnte das Landesarchiv Baden-Württemberg gewonnen werden, das – zum Beispiel durch die oben erwähnten Projekte »Piccard-Online« und »Bernstein« – über Sachkenntnis im inhaltlichen Umgang mit Wasserzeichen und umfassende informationstechnische Erfahrungen bei der Digitalisierung, Erfassung und Präsentation von Wasserzeichen verfügt. Am Projektvorhaben sind darüber hinaus die Bayerische Staats-

[1] Handschriftenerschließungszentren befinden sich zur Zeit an folgenden Bibliotheken: Staatsbibliothek zu Berlin, Universitätsbibliothek Johann Christian Senckenberg Frankfurt am Main, Universitätsbibliothek Leipzig, Bayerische Staatsbibliothek München, Württembergische Landesbibliothek Stuttgart und Herzog August Bibliothek Wolfenbüttel.

bibliothek München und die Universitätsbibliothek Leipzig beteiligt, die wie die Württembergische Landesbibliothek ihre Wasserzeichensammlungen digitalisieren und in das geplante System einbringen möchten.

Die in der Bayerischen Staatsbibliothek vorliegenden Durchzeichnungen und Durchreibungen von Wasserzeichen sind über einen längeren Zeitraum in verschiedenen Projekten der Handschriftenabteilung beziehungsweise des DFG-geförderten Handschriftenerschließungszentrums erstellt worden. Insgesamt beläuft sich die Zahl der derzeit vorhandenen Wasserzeichennachweise auf circa 4.900.

In der Universitätsbibliothek Leipzig wird seit 2004 kontinuierlich eine Wasserzeichenkartei aufgebaut, um die Beleglage für Wasserzeichen aus in Ostdeutschland verwendeten Papieren zu verbessern. Derzeit umfasst die Leipziger Sammlung circa 2.200 Nachweise; durch ein laufendes Projekt zur Erschließung deutschsprachiger mittelalterlicher Handschriften der Sächsischen Landesbibliothek – Staats- und Universitätsbibliothek Dresden wird sie bis Mitte 2010 auf circa 2.800 Belege anwachsen.

Bereits seit 2001 arbeitet das Landesarchiv Baden-Württemberg im Bereich der wissenschaftlichen Erschließung von Wasserzeichen aus mittelalterlichen Handschriften und Archivalien eng mit der Kommission für Schrift- und Buchwesen des Mittelalters der Österreichischen Akademie für Wissenschaften zusammen. Wesentlich intensiviert wurde diese Kooperation im Rahmen des DFG-Projektes »Piccard-Online« und des EU-Projektes »Bernstein«. Die Kommission konnte nun als Kooperationspartner für das geplante Projekt gewonnen werden und soll insbesondere am technischen Aufbau des Informationssystems und am Abgleich von »Piccard-Online« mit den gedruckten Piccard-Bänden mitwirken.

2 Zusammenfassung des Projektvorhabens

Ziel des geplanten Projekts ist der Aufbau eines gemeinsamen, Client-Server-basierten Informationssystems für Wasserzeichen und deren Beschreibungen für die DFG-Handschriftenzentren, bestehend aus jeweils einer Komponente für die dezentrale Dateneingabe, die zentrale Datenverwaltung und die homogene Online-Präsentation dieser Daten. Ein solches System soll den Aufbau und die Verwaltung digitaler Wasserzeichensammlungen erleichtern, so dass heterogene Wasserzeichenbestände deutscher Bibliotheken und Archive erstmals über einen zentralen Zugriffspunkt im Internet für die Wissenschaft, etwa in den Bereichen Wasserzeichenforschung und Handschriftenerschließung, nutzbar gemacht würden.

Im Gegensatz zum Bernstein-Portal, das vor allem eine Metasuche für die On-The-Fly-Abfrage dezentral gehosteter Datenbanken bietet und aufgrund der äußerst heterogenen Datenbasis lediglich eingeschränkte Funktionalitäten bereitstellen kann, soll das geplante Wasserzeichen-Informationssystem sämtliche Daten in einer zentralen

Datenbank bereithalten, was gegenüber der Abfrage in verteilten Datenpools auch schnellere Antwortzeiten ermöglicht. Durch die einheitliche Verwendung desselben Erfassungstools sowie durch die Festlegung von Erschließungsrichtlinien und einer gemeinsamen Klassifikation, die ebenfalls im Projekt erarbeitet werden sollen, lassen sich außerdem differenziertere Recherchen mit genaueren, strukturierten Ergebnissen einschließlich einer geografischen Visualisierung ermöglichen. Darüber hinaus soll das Wasserzeichen-Informationssystem als nationaler Aggregator für das europäische Bernstein-Portal fungieren.

Nach dem erfolgreichen Modell der Einbanddatenbank soll das System gleichzeitig eine gemeinsame Plattform für die digitale Aufbereitung vorhandener Wasserzeichensammlungen und ein kooperatives Arbeitsinstrument für die verteilte Erfassung neuer Wasserzeichennachweise aus der laufenden Handschriften- und Nachlasskatalogisierung darstellen. Abweichend vom Modell der Einbanddatenbank ist im beantragten Projekt eine redaktionelle Betreuung vorgesehen.

Nachfolgend werden die Arbeitsbereiche des Vorhabens im Einzelnen vorgestellt.

3 Aufbau der technischen Infrastruktur

Die Bereitstellung der technischen Infrastruktur für die dezentrale Dateneingabe und die zentrale Datenhaltung und -präsentation soll vorwiegend vom Landesarchiv Baden-Württemberg und der Österreichischen Akademie der Wissenschaften geleistet werden. Dazu zählt die Einrichtung der Systemarchitektur, mit anderen Worten die Implementierung der einzelnen Komponenten (das heißt des Erfassungsmoduls als Terminalserver-Installation, der Produktionsdatenbank, der Präsentationsdatenbank mit Präsentationsmodul) und das Einrichten der dezentralen Einzelarbeitsplätze als Terminalserver-Clients. Des Weiteren sollen Vernetzungsmöglichkeiten mit fachverwandten Systemen und Portalen geprüft und, wenn möglich, umgesetzt werden, um den Nutzwert des Informationssystems für die Forschung und Handschriftenerschließung weiter zu erhöhen. Darüber hinaus sollen Verknüpfungsmöglichkeiten mit dem Informationssystem »Manuscripta Mediaevalia« evaluiert und gegebenenfalls im Rahmen einer Schnittstelle realisiert werden.

3.1 Erfassungsmodul

Als eine der Hauptkomponenten des Informationssystems soll zunächst ein Erfassungsmodul (Produktionsdatenbank) für die dezentrale Dateneingabe und -verwaltung geschaffen werden. In dieses Backend sollen zum einen bereits bestehende Sammlungen wie zum Beispiel »Piccard-Online« integriert werden können, zum anderen soll den DFG-Handschriftenzentren ein Tool zur Verfügung gestellt werden, mit dem neue Wasserzeichensammlungen dezentral erschlossen werden können. Sämtliche Daten (Bilddaten und Beschreibungen) könnten anschließend unter einer einzigen Oberfläche verwaltet werden und würden in einer einheitlichen Datenstruktur vorliegen.

Die Grundlage für das Modul kann ein bereits bestehendes Tool bieten, das von Dr. Viktor Karnaukhov, Russische Akademie der Wissenschaften, für die Österreichische Akademie der Wissenschaften programmiert wurde und dort seit 1997 für die Beschreibung von Wasserzeichen verwendet wird (»Watermark Processing and Database Management Toolkit«, kurz »Watermark Toolkit«). Diese Software basiert auf der Programmiersprache C++ und läuft unter dem Betriebssystem Windows XP oder höher. Über die Benutzeroberfläche können Metadaten zu einzelnen Wasserzeichennachweisen (zum Beispiel Referenznummer, Motiv, Datierung, Fundort, Abmessungen) eingetragen und anschließend mit Digitalisaten verknüpft werden. Bilddaten können aus der Software heraus erzeugt oder nachträglich eingebunden werden und werden in einer lokalen Datenbank abgelegt. Als Grundlage für die Erschließung dient eine hierarchische Klassifikation, die über ein Graphical User Interface leicht verwaltet werden kann und die motivische Einordnung nicht vorhandener Nachweise komfortabel ermöglicht. Das Tool unterstützt außerdem die halbautomatische Vermessung der digitalisierten Wasserzeichen, indem per Maus ein Rechteck um das Papierzeichen gezogen wird. Ebenso wird der Vergleich und die Identifizierung erfasster Wasserzeichen über Motiv, Höhe und Breite ermöglicht. Zur Verifizierung einer vermuteten Identität können die Zeichen transparent übereinandergelegt werden.

Es ist beabsichtigt, das »Watermark Toolkit« im Zuge des Projektvorhabens an die Bedürfnisse der deutschen Handschriftenzentren anzupassen und mit einer zentralen MySQL-Datenbank für die Datenhaltung zu verbinden. Modifikationen müssten insbesondere im Hinblick auf die geplante dezentrale Datenerfassung und zentrale Datenverwaltung und die damit einhergehende erforderliche umfassende Benutzer- und Rechteverwaltung erfolgen. Die Weiterentwicklung der Software ist durch Viktor Karnaukhov in Zusammenarbeit mit der Österreichischen Akademie der Wissenschaften geplant und soll durchgehend vom Landesarchiv Baden-Württemberg begleitet werden, das auch das fertige Tool in die Gesamtarchitektur des Informationssystems integrieren wird.

3.2 Präsentationsmodul

Für die Darstellung der Inhalte im Internet ist die Entwicklung eines Online-Präsentationssystems vorgesehen. Dieses Frontend soll Wissenschaftlern und Handschriftenkatalogisierern einen Webzugriff auf sämtliche im Rahmen des Projekts erschlossenen Wasserzeichensammlungen bieten. Das Präsentationsmodul könnte auf der bereits erprobten Oberfläche von »Piccard-Online« aufbauen, das für die Verwendung als Komponente des Wasserzeichen-Informationssystems weiterentwickelt und mit zusätzlichen Funktionalitäten wie einer geografischen Visualisierung versehen werden soll.

Die Präsentationsdatenbank soll auf Grundlage der oben erwähnten MySQL-Datenbank des Erfassungsmoduls modelliert und über eine Schnittstelle mit den geprüften und freigegebenen Daten und digitalen Reproduktionen der Produktionsdatenbank gefüllt werden. Hierbei ist die Integration zusätzlicher Datenelemente, ein Datenmapping zur Optimierung der Performanz und eine partielle Denormalisierung der Daten notwendig. Auf Basis der Datenbank kann anschließend die Entwicklung des Präsentationsmoduls erfolgen, wobei auch eine Komponente für den unmittelbaren Vergleich digitalisierter Wasserzeichen unter Verwendung der Layer-Technologie, die Integration der Suchmaschine Lucene für die Volltextrecherche und die Implementierung eines Geografischen Informationssystems (GIS) vorgesehen sind.

Das Präsentationssystem soll vom Landesarchiv Baden-Württemberg konzipiert und umgesetzt werden.

3.3 Kartografische Visualisierung von Erschließungsergebnissen

Die Implementierung eines Geografischen Informationssystems könnte dem Nutzer einen geografischen Zugang zu den erschlossenen Wasserzeichen bieten und räumliche Zusammenhänge leichter erkennbar machen. Dazu ist beabsichtigt, bei der Beschreibung der Wasserzeichen im Erfassungsmodul Angaben zum Aufbewahrungsort des Stückes und zum Beschreibort des Papiers aufzunehmen, soweit Informationen dazu vorliegen. Zugehörige Ortsdaten (unter anderem Ortsname, einheitlicher Schlüssel, Geokoordinaten) könnten recherchiert und mit den Wasserzeichendatensätzen verknüpft werden, die Ausgabe dieser Informationen dann mittels eines Web-GIS im Rahmen der Online-Präsentation erfolgen. Auf diese Weise könnten Verbreitungswege und -gebiete von Papier und Wasserzeichenmotiven sichtbar gemacht werden.

4 Aufbereitung von Inhalten für das Wasserzeichen-Informationssystem

Neben der technischen Entwicklung des Informationssystems bildet die Anreicherung mit Inhalten, das heißt die Digitalisierung und Erschließung von Wasserzeichennachweisen, den zweiten Schwerpunkt des Projektvorhabens.

So sollen etwa 12.000 neue Wasserzeichenbelege von der Württembergischen Landesbibliothek Stuttgart, der Bayerischen Staatsbibliothek München und der Universitätsbibliothek Leipzig eingebunden werden, indem sie vor Ort dezentral im Erfassungsmodul beschrieben und mit digitalisierten Bilddaten verknüpft werden. Hierbei ist zu beachten, dass die Beschaffenheit und das Alter der Unterlagen aus den einzelnen Beständen und Sammlungen stark variiert. Um dennoch eine einheitliche Dateneingabe zu gewährleisten, ist vorgesehen, Richtlinien für die Beschreibung von Wasserzeichen festzulegen und die Dateneingabe an zentraler Stelle im Landesarchiv und der Landesbibliothek in Stuttgart zu koordinieren und redaktionell zu betreuen. Anhand dieses Vorgehens kann der Nutzwert des Präsentationssystems wesentlich gesteigert werden, da die Erschließungsdaten sehr strukturiert dargestellt und differenzierte Suchfunktionen mit besseren Ergebnissen angeboten werden können.

4.1 Dezentrale Digitalisierung und Erschließung der Wasserzeichennachweise

Die einzelnen Wasserzeichendurchzeichnungen und -durchreibungen, die in der geplanten Projektphase zu bearbeiten sein werden, sind auf Einzelblättern festgehalten und sollen als Erstes mit einer Auflösung von 300 dpi – bezogen auf die Originalvorlagengröße – digitalisiert und jeder Scan in einer vorgegebenen Verzeichnisstruktur mit festgelegtem Dateibenennungsschema gespeichert werden. Aus den gelieferten Bilddateien können später die einzelnen Wasserzeichenmotive mittels Bildbearbeitung händisch oder – soweit möglich – automatisiert isoliert und abgespeichert werden. Die dann folgende Erschließung der Wasserzeichen soll in den beteiligten Handschriftenzentren in Stuttgart, München und Leipzig mithilfe des oben beschriebenen Erfassungsmoduls durchgeführt werden. Diese Dateneingabe muss mit einer Recherche nach verwandten Nachweisen im bereits vorhandenen Datenpool und deren Verknüpfung verbunden sein und unter Berücksichtigung einheitlicher Erschließungsgrundsätze erfolgen.

4.2 Richtlinien für die Erschließung und Anpassung der Klassifikation

Um trotz der verschiedenartigen Ausgangsmaterialien eine homogene Eingabe in das Erfassungsmodul zu erreichen, ist die Entwicklung und Festlegung differenzierter Erschließungsrichtlinien vorgesehen. Die Erfassungsmaske sollte entsprechend einen zweiteiligen Felderpool mit Pflicht- und Optionsfeldern bieten, mit dem die neu zu erfassenden Sammlungen aufgenommen werden können. Grundlage für die Richtlinien bildet ein hierarchisch strukturiertes Klassifikationsmodell, welches im Rahmen des Bernstein-Projektes auf der Basis der bestehenden großen Wasserzeichensammlungen »Piccard-Online«, »Wasserzeichen des Mittelalters« und »Watermarks in Incunabula printed in the Low Countries« und unter Berücksichtigung des Wasserzeicheninventars von Briquet entwickelt wurde und auf der obersten Ebene zwölf verschiedene Motivgruppen vorsieht (Frauenknecht et al.) Diese werden weiter differenziert; innerhalb der einzelnen Ebenen können bis zu 999 Klassifizierungspunkte angelegt werden. Es ist beabsichtigt, die bestehenden drei Motivebenen im Zuge der Erschließungsarbeiten systematisch auf insgesamt bis zu zehn Ebenen zu erweitern.

4.3 Redaktion und Koordination der Dateneingabe

Da die Dateneingabe dezentral erfolgen soll, ist eine redaktionelle Pflege der Daten sowie eine Evaluation der Erfassung erforderlich. Hierfür ist die Einrichtung einer Redaktions- und Koordinationsstelle geplant, deren Aufgaben während der Projektlaufzeit vom Landesarchiv Baden-Württemberg und der Württembergischen Landesbibliothek gemeinsam wahrgenommen werden sollen. In ihren Zuständigkeitsbereich fällt auch die Vorbereitung des multilingualen Zugriffs auf die Datenbank und die Referenzierung auf bestehende Wasserzeichensammlungen: Als Nachnutzung aus dem laufenden Bernstein-Projekt wird ein gemeinsames Vokabular zur Verfügung gestellt, mit dem die unterschiedliche Kongruenz bestehender Wasserzeichensammlungen ausgeglichen und gleichzeitig ein multilingualer Zugriff ermöglicht werden kann.

Des Weiteren bildet die Referenzierung auf einschlägige, bereits veröffentlichte Wasserzeichensammlungen (gedruckt oder online) eine entscheidende Komponente, um Wasserzeichenbelege in den Kontext der Wasserzeichen- und Handschriftenforschung einzubinden. Die neu zu erfassenden Wasserzeichensammlungen sollen auf die maßgeblichen Wasserzeichendatenbanken im Portal »Bernstein« referenzieren. Es ist beabsichtigt, zunächst bestehende Verweise zu prüfen oder neue zu erstellen und nach der endgültigen Überprüfung durch die Redaktionsstelle dann als Link anzulegen. Ebenfalls soll mittels eines Links auf die »Manuscripta Mediaevalia« und durch Referenzen auf die gedruckten Findbücher von Piccard sowie eventuell auf Briquet verwiesen werden, dessen Digitalisierung gerade von der Kommission für Schrift- und Buchwesen des Mittelalters in Wien vorbereitet wird.

4.4 Einbindung von »Piccard-Online« und Abgleich mit Piccard-Findbüchern

Neben der Erschließung neuer Wasserzeichen ist auch die Integration der ca. 92.000 Datensätze der »Piccard-Online«-Anwendung und der bislang noch nicht digitalisierten Wasserzeichen der gedruckten Piccard-Bände 3-17 vorgesehen. Beide basieren auf der für die Wasserzeichenbeschreibung einschlägigen Referenzsammlung Piccard.

Um »Piccard-Online« sinnvoll in das neue Wasserzeichen-Informationssystem integrieren zu können, soll die bisherige Piccard-Systematik von Bearbeitern im Landesarchiv Baden-Württemberg auf die neue, oben beschriebene Klassifikation gemappt und diese Änderungen in der »Piccard-Online«-Datenbank umgesetzt werden.

Da bei der Drucklegung der Findbücher eine nicht genau verifizierbare Zahl an Karten ausgeschieden wurde, besteht eine Diskrepanz zwischen »Piccard-Online« und den Piccard-Findbüchern. Um eine Online-Suche im gesamten Datenbestand Piccard zu ermöglichen, sollen die abschriftlich vorliegenden Metadaten zu den Papiermarken der Findbücher in eine mit »Piccard-Online« kompatible Form gebracht werden. Darüber hinaus soll für einige ausgewählte, häufig verwendete Motive ein Abgleich der Marken in »Piccard-Online« und den Findbüchern mit dem Ziel der Identifizierung identischer Marken durchgeführt werden. Die Integration der Piccard-Bände und der Abgleich mit »Piccard-Online«, für die sich die Kommission für Schrift- und Buchwesen des Mittelalters aufgrund von langjährigen Vorarbeiten bereitwillig angeboten hat, könnte die Referenzierung neu registrierter Wasserzeichen erleichtern und beschleunigen.

5 Ausblick

Mit dem geplanten Wasserzeichen-Informationssystem soll erstmals ein Instrument geschaffen werden, das zwei Anliegen der Fachwelt zugleich erfüllt: Einerseits erhielten die Handschriftenzentren in Deutschland ein Werkzeug, mit dem sie selbständig neue digitale Wasserzeichensammlungen aufbauen und online zur Verfügung stellen könnten; andererseits hätten Wissenschaftler aus dem Bereich der Handschriftenbearbeitung und verwandten Forschungszweigen erstmalig die Chance, zentrale Recherchen und Vergleiche auszuführen und auf Ressourcen zuzugreifen, die bislang nicht oder nur auf verschiedene Zugangspunkte verteilt in digitaler Form nutzbar sind.

Zusätzlich zur Präsentation der Daten im Frontend des Informationssystems könnte eine Recherche und Einsichtnahme über externe Informationssysteme wie z.B. das Bernstein-Portal erfolgen, für welches das geplante Informationssystem als Aggregator fungieren soll.

Eine weitere Anreicherung der Datenbasis wäre auch nach Ablauf des geplanten Projektes jederzeit möglich und ist ausdrücklich erwünscht. Das Erfassungsmodul sollte sämtlichen DFG-Handschriftenzentren bereits nach der ersten Projektphase für den Gebrauch bereitgestellt werden und auch nach Projektabschluss weiterhin offen stehen, so dass diese stets neue digitale Wasserzeichensammlungen anlegen und einpflegen könnten. Das System soll auch nach Ende einer möglichen Förderung operativ weiterbetreut werden, wozu sich die Württembergische Landesbibliothek und das Landesarchiv Baden-Württemberg bereit erklärt haben.

Es wäre zu überlegen, ob nicht in einer späteren Förderungsphase gezielt weitere Sammlungen erschlossen werden sollen. Hier kämen beispielsweise die Nachweise aus neuzeitlichen Beständen der Hessischen Archive in Frage, die im Zuge der Katalogisierung von Musikhandschriften gesammelt wurden. Umfangreiche Wasserzeichensammlungen aus Inkunabeln sowie aus frühneuzeitlichen und modernen Papieren werden ferner auch im Deutschen Buch- und Schriftmuseum in Leipzig sowie in der Handschriftenabteilung der Staatsbibliothek zu Berlin - Preußischer Kulturbesitz aufbewahrt.

Weiterhin wäre vorstellbar, dass auch die Kommission für Schrift- und Buchwesen des Mittelalters ihre Inhalte in das Wasserzeichen-Informationssystem einpflegt. Die dortige Sammlung »Wasserzeichen des Mittelalters« beinhaltet derzeit 9.550 Wasserzeichen und wird jährlich um etwa 1.000 neue Marken erweitert – eine Anzahl neuer Nachweise, die den wissenschaftlichen Nutzen des Wasserzeichen-Informationssystems deutlich erhöhen könnte.

Wie sich an den genannten Beispielen erkennen lässt, sind zukünftigen Erweiterungen der Datenbasis keine Grenzen gesetzt. Zunächst stehen jedoch die Errichtung des Informationssystems sowie die Anreicherung mit Inhalten aus »Piccard-Online« und den drei beteiligten DFG-Handschriftenzentren im Vordergrund. Bereits damit wird der Wissenschaft ein wertvolles Werkzeug für die Erforschung von Wasserzeichen und Handschriften bereit gestellt.

Bibliographie

Bernstein – the memory of paper. <http://www.memoryofpaper.eu>.

Briquet, Charles Moïse. *Les Filigranes. Dictionnaire historique des marques du papier dès leur apparition vers 1282 jusq'en 1600.* 4 Bände. Amsterdam: The Paper Publ. Soc., 1968.

Einbanddatenbank. <http://www.hist-einband.de>.

Frauenknecht, Erwin, Peter Rückert und Maria Stieglecker. Bernstein Systematik. <http://www.bernstein.oeaw.ac.at/twiki/bin/viewfile/Main/DocumentsArchive?rev=1;filename=Bernstein_Systematik.pdf>.

International Database of Watermarks and Paper Use for Prints and Drawings c. 1450-1800. <http://www.wm-portal.net/niki/index.php>.

Lucene. <http://lucene.apache.org>.

Manuscripta Mediaevalia. <http://www.manuscripta-mediaevalia.de>.

Piccard, Gerhard, Hrsg. *Piccard Wasserzeichen: Veröffentlichungen der Staatlichen Archivverwaltung Baden-Württemberg. Die Wasserzeichenkartei Piccard im Hauptstaatsarchiv Stuttgart.* 17 Findbücher in 25 Bänden. Stuttgart: Kohlhammer 1961-1997.

Rückert, Peter, Jeannette Godau und Gerald Maier, Hrsg. *Piccard-Online. Digitale Präsentationen von Wasserzeichen und ihre Nutzung.* Stuttgart: Kohlhammer 2007.

Wasserzeichen des Mittelalters (WZMA). <http://www.ksbm.oeaw.ac.at/wz/wzma.php>.

Wasserzeichendatenbank des »Nederlands Interuniversitair Kunsthistorisch Instituut.« <http://www.iuoart.org> und <http://www.niki-florence.org>.

Watermark Terms. Vocabulary for watermark description. <http://www.bernstein.oeaw.ac.at/twiki/bin/viewfile/Main/DocumentsArchive?rev=1;filename=Watermark_Terms_v7_revised.pdf>.

Watermarks in Incunabula printed in the Low Countries (WILC). <http://watermark.kb.nl>.

Paläographie: Vom virtuellen Lernen zu neuen Perspektiven

Palaeography: From eLearning to New Research Horizons

Handschriften lesen lernen im digitalen Zeitalter

Silke Kamp

Zusammenfassung

Für das Erlernen der frühneuzeitlichen Kurrentschrift lässt sich der Anspruch des Tafelwerks, für nahezu jede Schreiberhand den passenden Leseschlüssel bereit zu halten, nur schwer mit den Lernzielen eines Hochschulseminars in Einklang bringen. Einführungsveranstaltungen zur Paläographie können nicht das Transkribieren Wort für Wort vermitteln, sie müssen vornehmlich auf das Sinnerschließen bauen. Hier bietet das Tafelwerk kaum Hilfestellung. Auch auf das individuelle Lernen und unterschiedliche Lerntypen sind Tafelwerke nicht ausgelegt. Weitere zeitgemäße Lernziele, wie kooperatives Lernen oder die Verbindung zu Forschung und Archiven bleiben, allein auf das Tafelwerk gestützt, ebenfalls außer Reichweite. Diese Lernziele lassen sich mit digitalen Ressourcen geradezu spielerisch verwirklichen. 1. Digitale Werkzeuge setzen bei der Vermittlung von Lesestrategien an: Ein Sütterlin-Font hilft Quellenbegriffe dem Schriftbild einer Kurrentschrift anzunähern (»Schlagwortsuche«), mit den Funktionen »Ausschneiden« und »Einfügen« lassen sich Leseschlüssel für bestimmte Schreiberhände erstellen (»individueller Leseschlüssel«). Auch kleine Lerneinheiten können so aus einzeln extrahierten Worten für das tägliche Üben entstehen, wie sie in David Postles' Lernprogramm »Early-modern Paleography« realisiert sind. Kombiniert mit einer weiteren Lesestrategie, dem »Lückentext«, lassen sich handschriftliche Quellen auch im Schulunterricht transkribieren. Animierte Schreibbewegungen helfen, sich ähnelnde Buchstaben zu unterscheiden. Diese Lesestrategien werden für Kurrentschrift derzeit am besten im Lernprogramm »Ad fontes« umgesetzt, Übungen zu einzelnen Buchstaben und zum Sinnerschließen bleiben jedoch Desiderate. 2. Digitalisierte Materialien vertiefen die Lernziele. Das Bereitstellen mehrerer digitalisierter Quellen einer Gattung vermittelt paläographische Kenntnisse, die sonst nur im Archiv gewonnen werden können. 3. Besondere Bedeutung kommt der Gestaltung der Lernplattform zu, wo neben der Bereitstellung der digitalen Werkzeuge und Materialien auch Zugang zu externen Ressourcen gewährt und in Arbeitsgruppen kooperatives Lernen initiiert werden kann.

Abstract

Albums of palaeographical plates claim to offer the key to reading the script of almost any scribe, but, in the case of teaching early-modern cursive, this claim is difficult to harmonize with the learning outcomes of a university seminar. Introductory palaeography courses cannot teach word-for-word transcription, but have to build upon the

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 111–122.

meaning of the text. For this, an album of plates offers little help. These albums are also not designed to accommodate private study and different learning styles. Further current educational objectives, such as cooperative learning or the interaction with research and archives, will remain beyond reach if their attainment relies upon such collections of plates. All these goals can be achieved easily by using digital resources. First, digital tools can be used to teach reading strategies: a Sütterlin font approximates catchwords to early-modern handwriting so that using copy-and-paste allows individuals to build their own key to particular handwritings. Small training units can then be created for daily practice, as has been demonstrated by David Postles' application "Early-modern palaeography". In combination with other reading strategies, such as filling-in blanks, transcribing manuscript sources can be made part of secondary eduction as well. Computer animations of writing motions helps students differentiate between similar letters. For cursive script, these strategies are currently best implemented by the application "Ad fontes". Exercises for individual letters and for understanding of the meaning of the text remain desiderata, however. Second, digitized material gives greater depth to the learning outcomes. The preparation of several different digitalized sources of a single genre teaches palaeographical skills that could otherwise only be gained in the archive. Third, the design of the learning platform is especially important because it provides not only digital tools and materials, but it also offers access to external resources and allows for cooperative studying in groups.

1 Vorbemerkungen

Web-basierte Seminare sind seit langem an den Hochschulen auf dem Vormarsch. Lernplattformen wie Blackboard oder Moodle haben bereits den klassischen Semesterapparat abgelöst, und Vorlesungen lassen sich statt im überfüllten Hörsaal bequem vor dem heimischen PC mit verfolgen. Als didaktisches Konzept hat sich das Blended-Learning durchgesetzt, da die Mischung von herkömmlicher Präsenzveranstaltung mit web-basiertem E-Learning Flexibilität und Begleitung des Lernens gleichermaßen realisiert (Geldsetzer und Strothmann). Studierende stehen dem Einsatz neuer Medien im Hochschulalltag nicht nur aufgeschlossen gegenüber, ihre Lernmotivation fällt in web-basierten Seminaren höher aus als in reinen Präsenzveranstaltungen (Mankel). Eine virtuelle Lernumgebung verbessert nicht nur das Lernklima, sie kommt insbesondere Studierenden mit Handicap entgegen. Sprechen darüber hinaus noch weitere Gründe für den Einsatz von Digitalisierungen? Dies soll hier am Beispiel der paläographischen Lehre diskutiert werden.

Diese Überlegungen basieren auf zwei Seminaren zur Paläographie der Frühen Neuzeit, die an der Universität Potsdam im Sommer- und Wintersemester 2008/2009 angeboten wurden. Im Mittelpunkt stand dabei die Vermittlung von Lesekompetenzen in

Kanzlei- und Kurrentschrift. Die Kurse sollten neben quellenbasierten Seminaren im Lehrangebot am Historischen Institut auch auf die Arbeit im Archiv vorbereiten. Daher fokussiert dieser Artikel auf den Nutzen digitaler Materialien und Werkzeuge für die Transkription frühneuzeitlicher Handschriften.

Zunächst soll geklärt werden, welche Anforderungen sich aus diesen Lernzielen für die Verwendung digitaler Ressourcen in der Lehre ergeben und warum diese mit dem Tafelwerk nur eingeschränkt erreicht werden können. Daran schließt sich eine Diskussion der Voraussetzungen an, die für den erfolgreichen Einsatz digitaler Ressourcen erfüllt sein müssen. Zum Schluss sollen Möglichkeiten und Perspektiven für die Verwendung digitaler Materialien und Werkzeuge in der paläographischen Lehre aufgezeigt werden.

2 Anforderungen

Ganz gleich, ob ein Kurs zur Paläographie als Einführung für quellenbasierte Seminare konzipiert ist oder auf die Archivarbeit vorbereiten soll, im Laufe eines Semesters können nur die Grundlagen für das Transkribieren von Kanzlei- und Kurrentschrift gelegt werden. Nur ein geringer Teil der Studierenden erwirbt ausreichende Lesekompetenzen, um Dokumente lückenlos zu transkribieren. Des Weiteren kommt zwar der eigenen Transkriptionserfahrung bei der Vermittlung paläographischer Kenntnisse ein hoher Stellenwert zu, doch lassen sich Lesestrategien nicht ohne weiteres in didaktische Konzepte übersetzen. Was einem selbst beim Entziffern nützlich und einleuchtend erscheint, ist nicht unbedingt anschaulich und daher überzeugend im Seminar zu vermitteln. Genau hierin liegen die Schwierigkeiten beim Einsatz des Tafelwerks.

Ein Standardwerk der Paläographie ist Paul Arnold Gruns Leseschlüssel (Grun). In Schrifttafeln werden gängige Varianten der einzelnen Groß- und Kleinbuchstaben in Kanzlei- und Kurrentschrift wiedergegeben. Mit ihrer Hilfe lassen sich Handschriften vom 16. bis zum 19. Jahrhundert entziffern. Für das Erlernen der um 1900 entwickelten Sütterlinschrift bietet sich hingegen das Lehrbuch von Harald Süß, Deutsche Schreibschrift an (Süß). Neben dem Entziffern steht hier das Schreiben der deutschen Schrift im Vordergrund. Für das Transkribieren frühneuzeitlicher Handschriften bietet das Sütterlin freilich einen Zugang, wenn auch einen sehr aufwendigen. Auch Schrifttafeln wie die von Grun sind zweifelsohne als Nachschlagewerk ein unerlässliches Hilfsmittel. Doch wenn es um die Vermittlung von Lesekompetenzen geht, geben weder Tafelwerk noch Lehrbuch realistische Lernziele vor, die innerhalb eines Semesters erreicht werden können: Das Lesen der Kurrent- oder Kanzleischrift kann innerhalb eines Kurses kaum erlernt werden, auch nicht über das Schreibenlernen des Sütterlins.

Das bedeutet für die Lehre, die Kursteilnehmer von Anfang an auf die Schlüsselstellen eines Dokumentes aufmerksam zu machen. Von diesen ausgehend, müssen sie den

Sinn des Schriftstückes erfassen lernen, anstatt es auf Anhieb Wort für Wort transkribieren zu wollen. Doch dies entspricht genau dem, wie Studierende bei der Transkription vorgehen. Dieses Verhalten wird durch Tafelwerke noch verstärkt, da Schrifttafeln durch ihre Fülle an Buchstabenvarianten suggerieren, selbst für schwierigste Lesungen den passenden Schlüssel bereit zu halten. Die Arbeit mit dem Tafelwerk verführt also dazu, Handschriften Wort für Wort zu transkribieren und den Sinn eines Schriftstückes außer Acht zu lassen.

Eine weitere Schwäche des Tafelwerks ist, dass es den visuellen Lerntyp bevorzugt. Dies wird insbesondere bei der Unterscheidung von ähnlich aussehenden Buchstaben problematisch. Tafelwerke nehmen zwar Rücksicht auf die gängigen Varianten einzelner Buchstaben, aber sie erklären nicht, wie man z.B. h und s voneinander unterscheiden kann. Selbst wenn, wie im Übungsbuch von Süß, auf die Ähnlichkeit dieser Buchstaben verwiesen wird, entwickeln Studierende hierfür nur selten ein Problembewusstsein. Der haptische Lerntyp kann erst aus dem Wissen, wie diese Buchstaben ausgeführt werden, zu ihrer Unterscheidung gelangen. Wenn er das Schreiben nicht nur einzelner Buchstaben sondern des Sütterlins allgemein beherrscht, kann er Kanzlei- und Kurrentschrift leichter entziffern. Doch muss das Schreiben der deutschen Schrift erst ebenso mühevoll erlernt werden wie das Lesen. Auch das Üben einzelner Buchstaben wird durch das Tafelwerk nicht unterstützt und das Schreibenlernen des Sütterlins stellt für das Seminar keine sinnvolle Alternative dar.

Die Lernziele Sinnerschließen und das Üben einzelner Buchstaben lassen sich demnach auf Tafelwerk und Übungsbuch gestützt kaum vermitteln. Andere Lernziele, wie das kooperative Lernen, das Einbeziehen aktueller Forschungsprojekte und der Kontakt zu Archiven erfordern einen Medienwechsel. Alle diese Lernziele können durch den Einsatz digitaler Ressourcen geradezu spielerisch erreicht werden, wenn folgende Voraussetzungen gegeben sind:

3 Voraussetzungen

Dem Tafelwerk sind in der paläographischen Lehre wie gezeigt Grenzen gesetzt. Damit diese mit Hilfe digitaler Ressourcen überschritten werden können, müssen drei Dinge beachtet werden: die stufenweise Aufbereitung digitaler Materialien, die Vermittlung von Medienkompetenzen zum Umgang mit digitalen Werkzeugen und die angemessene Gestaltung der virtuellen Lernumgebung.

Handschriftliche Quellen müssen in digitalisierter Form vorliegen, um sie mit digitalen Werkzeugen bearbeiten zu können. Wenn Quellen aus der eigenen Forschungsarbeit im Seminar benutzt werden sollen, stehen nicht immer Faksimiles in digitalisierter Form aus den Archiven zur Verfügung. Oftmals muss der Dozent die Digitalisierung anhand von Xerox-Kopien selbst vornehmen und das Gescannte gegebenenfalls mit

Hilfe von Bildbearbeitungsprogrammen aufbereiten (Kontrast erhöhen). Für Faksimiles ist eine gute Auflösung wichtig, damit einerseits Ausschnitte beliebig vergrößert werden können, ohne dass die Lesbarkeit darunter leidet. So liefern andererseits Ausdrucke selbst auf älteren Druckermodellen noch ansprechende Druckergebnisse. Studierende greifen neben der Digitalisierung gern auf den Papierausdruck zurück. Hier ändert sich das Lernverhalten nicht, es wird allenfalls ergänzt. Diese Beobachtung konnte bereits für das E-Learning gemacht werden (Mankel).

Weder kann vorausgesetzt werden, dass die Seminarteilnehmer über die gleichen technischen Voraussetzungen verfügen, um digitale Ressourcen zu nutzen, noch kann von einer einheitlichen Medienkompetenz ausgegangen werden. Will man sich die höhere Motivation der Studierenden zu web-basiertem Lernen zu Nutze machen, sollte das Seminar zu Beginn auch eine Einführung geben, die zum Arbeiten mit der virtuellen Lernumgebung befähigt (Mankel).

Der Vermittlung entsprechender Medienkompetenzen kommt deswegen besondere Bedeutung zu, weil die virtuelle Lernumgebung für den Einsatz digitaler Ressourcen zentral ist. Ihr Erfolg hängt maßgeblich von ihrer übersichtlichen Struktur ab. Die Lernplattform stellt nicht nur die digitalisierten Materialien bereit, sondern auch die digitalen Werkzeuge für das individuelle und kooperative Bearbeiten der Quellen. Letzteres lässt sich initiieren, indem auf der Lernplattform Arbeitsgruppen eingerichtet werden. Ihnen können bestimmte Aufgaben zugewiesen werden, die in der Gruppe zu lösen sind. Bei der Gestaltung der Lernplattform gilt es ferner zu berücksichtigen, dass nicht jeder Studierende über einen uneingeschränkten Internetzugang (Flatrate) verfügt. Dies machen Erfahrungsberichte zum Studium an der Universität Potsdam deutlich (»Ich fürchte, wir haben ein Problem«). Die Lernplattform selbst sollte daher in der Bereitstellung digitaler Werkzeuge wiederum gemischt (blended) sein, und zwar was Online-Werkzeuge anbelangt und Downloads. Bei den Downloads gilt es, auf die Dateigröße zu achten. Eine Dateigröße bis zwei MB wird meiner Erfahrung nach von den meisten Studierenden akzeptiert. Gegebenenfalls empfiehlt es sich, zu Seminarbeginn die technischen Möglichkeiten der Kursteilnehmer abzufragen und das Lernmaterial darauf abzustimmen. Für Faksimiles sind neben dem komprimierten PDF-Format JPGs in verschiedenen Auflösungsgraden sinnvoll, um Auflösung und Verfügbarkeit digitaler Materialien für die Nutzer der Lernplattform zu optimieren. Diese Faksimiles lassen sich in einzelnen Lektionen mit steigendem Schwierigkeitsgrad organisieren. Transkriptionsaufgaben, die eine direkte Eingabekontrolle gewähren, stellen eine sinnvolle Ergänzung dar. Sie lassen sich auf der Lernplattform etwa als Quiz realisieren, indem die richtige Transkription einzelner Worte oder Zeilen abgefragt wird. Die direkte Rückmeldung steigert Lernerfolg und Lernmotivation gleichermaßen. Auch das Quiz lässt sich nach steigendem Schwierigkeitsgrad anschaulich strukturieren. Quizfragen bieten darüber hinaus noch weitere Vorteile: Sie können unabhängig voneinander beantwortet werden und sie entsprechen kleinen Lerneinheiten. Dies ermöglicht einerseits auch Studierenden ohne eigenen Internetzugang diese Aufgaben in Bibliotheken, Computer-Pools oder Internetcafés zu lösen. Andererseits sind kleine Lerneinheiten auch deshalb ratsam, weil das regelmäßige Arbeiten mit Handschriften den größten Lernerfolg verspricht. Als optimal gelten tägliche Übungseinheiten von 15 bis 20 Minuten. Dieses Lernverhalten kann durch die virtuelle Lernumgebung noch unterstützt werden. Über externe Ressourcen kann insbesondere der Kontakt zu Archiven und damit die Verbindung von Lehre und Forschung hergestellt werden. Verweise zu digitalen Editionsprojekten oder universitären Angeboten zu Hilfswissenschaften reichern die Lernplattform zusätzlich an und halten sie aktuell und flexibel. Sind diese Voraussetzungen erfüllt, ergeben sich beim Einsatz digitaler Ressourcen in der paläographischen Lehre Möglichkeiten, die weit über die des Tafelwerks hinausreichen.

4 Möglichkeiten

Wie können die an die paläographische Lehre im digitalen Zeitalter gestellten Anforderungen erfüllt werden? Hier sind in erster Linie die Lesestrategien zu benennen, die sich auf das Tafelwerk gestützt nur mühsam vermitteln lassen. Ich habe diese Strategien aus meiner eigenen Erfahrung mit handschriftlichen Quellen abgeleitet und sie der »Lückentext«, die »Schlagwortsuche«, der »individuelle Leseschlüssel« und die »unleserliche Hand« genannt.

Die Methode »Lückentext« geht davon aus, dass zuerst einzelne Buchstaben erkannt werden, aus denen sich nach und nach die Worte zusammensetzen. Dabei wird zuerst die Anzahl der Buchstaben eines Wortes bestimmt und durch Platzhalter (Punkte) angedeutet. Nach und nach werden die Punkte durch Buchstaben ersetzt und die Lücken allmählich durch Erschließen von Zusammenhängen gefüllt. Hier macht man sich auch den Umstand zu Nutze, dass innerhalb einer Handschrift Buchstaben einmal leserlicher und ein anderes Mal unleserlicher ausgeführt vorliegen, bedingt auch durch vom nächsten Buchstaben abhängige Ligaturen und Verschränkungen. Die Methode Lückentext trägt auch der Beobachtung Rechnung, dass Studierende individuelle »Vorlieben« für einzelne Buchstaben ausbilden, jeder also bestimmte Buchstaben leichter entziffern kann als andere. Umgekehrt sind auch die Schwierigkeiten beim Erkennen einzelner Buchstaben individuell verteilt. Gerade auf diese Stärken und Schwächen kann das Tafelwerk keine Rücksicht nehmen.

Die Methode »Schlagwortsuche« basiert auf dem Vergleich von Mustern. Für Anfänger im Kurrent- oder Kanzleischriftlesen ist es leichter, sich statt auf den ganzen Text auf einzelne Worte zu konzentrieren, und sich, wie ein funktionaler Analphabet, nach dem Schriftbild zu orientieren. Ein Dokument wird daraufhin geprüft, ob bestimmte (Schlag-)Wörter in ihm auftauchen oder nicht. Um das Auffinden der Schlagworte zu erleichtern, werden sie in alphabetischer Reihenfolge dem Schriftbild der Quelle angeglichen. Der Erfolg dieser Methode hängt davon ab, dem Schriftbild der Quelle möglichst nahe zu kommen. Auch hierbei bietet das Tafelwerk wenig Hilfestellung, da Wortbeispiele selten mit dem Vokabular von Quellentexten übereinstimmen. So ist man gezwungen, die Quellen nach den Wortbeispielen im Tafelwerk auszusuchen.

Der »individuelle Leseschlüssel« ist eine Erweiterung des Tafelwerks. Bei sehr individuellen Handschriften oder einer bestimmten, häufig wiederkehrenden Schreiberhand ist das Erstellen eines individuellen Leseschlüssels für die Entzifferung ganzer Konvolute eine Arbeitserleichterung. Zugleich kann auch im Sinne einer Schriftvergleichung ein möglicher Wechsel der Schreiberhände erkannt werden. Der Nachteil dieser Methode ist, dass der individuelle Leseschlüssel die einzelnen Buchstaben möglichst präzise in der Ausführung des Schreibers wiedergeben muss. Inwieweit dies gelingt, hängt also von den künstlerischen Fähigkeiten des Anwenders ab.

Unter der Methode »unleserliche Hand« sind eine Reihe von Tipps zusammengefasst, die das Entziffern einer schwer lesbaren Textpassage – egal ob der Eigenart des Schreibers oder dem Zustand des Dokuments geschuldet – erleichtern sollen. Wiederum steht nicht die vollständige Transkription im Vordergrund, sondern das Erkennen wichtiger Textstellen. Zunächst geht es darum, einzelne Sätze zu erkennen, anschließend in den Sätzen das Verb zu finden und schließlich wichtige Substantive und handelnde Personen, Orte und Zeitangaben aufzuspüren. Anders als bei der Schlagwortsuche wird hier also nach unbekannten Worten gesucht. Sind die essentiellen Textstellen identifiziert, geht es erst im zweiten Schritt um deren Entzifferung. Dieses Sinnerschließen wird durch die Eigenschaft des Tafelwerks, wie gesehen, nicht unterstützt, um nicht zu sagen verstellt.

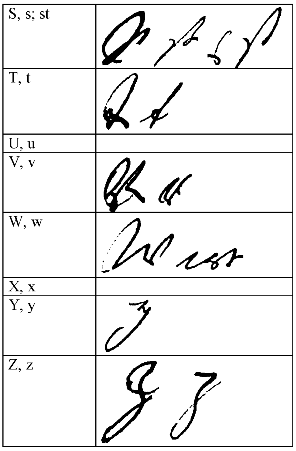
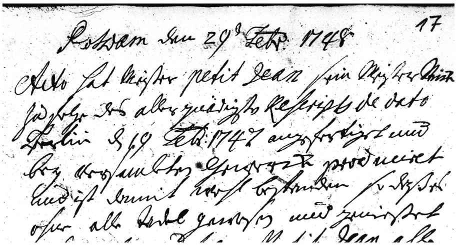

S, s; st
T, t
U, u
V, v
W, w
X, x
Y, y
Z, z

Abbildung 1. Beispiel für einen individuellen Leseschlüssel.



Potsdam den 29t Febr: 1748 17

Acto hat Meister petit Jean sein Meister [illegible]

zufolge des allergnädigsten Rescripts de dato

Berlin d. 19 Febr:1747 ausgefertiget und

bey versamleten Gewerck produciret

und ist damit recht befunden [illegible]

ohne alle Tadel [illegible] und [illegible]

Abbildung 2. Quelle für den individuellen Leseschlüssel.

Für das Erlernen dieser vier Lesestrategien steht eine Reihe von digitalen Werkzeugen bereit. Für die Schlagwortsuche bietet sich als digitales Werkzeug die Schriftart »Sütterlin« an. Sie ist Bestandteil des in Kooperation der Universitäten Potsdam und des Saarlandes entwickelten Sütterlin-Lernprogramms SLP 2000, mit dessen Hilfe das Lesen der von Ludwig Sütterlin entworfenen Schreibschrift erlernt werden kann. Im Rahmen des Programms steht sie im Internet zum Downloaden zur Verfügung und lässt sich problemlos auf PC oder Mac installieren. Mit Hilfe dieses Fonts lassen sich ohne großen Aufwand Schlagwortlisten erstellen, mit denen insbesondere handschriftliche Quellen vom beginnenden 19. Jahrhundert bearbeitet werden können. Der Nutzen der Sütterlinschrift für das Erlernen der Kurrentschrift wird jedoch von den Studierenden unterschiedlich erlebt. Auch für das Erstellen von individuellen Leseschlüsseln bieten digitale Werkzeuge eine große Erleichterung. Mit Hilfe von Bildbearbeitungsprogrammen lassen sich über die Ausschneiden- und Einfügen-Funktionen einzelne Buchstaben tabellarisch zu einem Leseschlüssel für eine bestimmte Schreiberhand zusammenfügen. Mit Bildbearbeitungs- und Präsentationsprogrammen lässt sich die Methode »unleserliche Hand« dahingehend unterstützen, dass Schlüsselstellen eines Dokumentes graphisch hervorgehoben werden können.

Für das Üben einzelner Buchstaben stellen Bildbearbeitungsprogramme wiederum ein nützliches Werkzeug dar, um Übungen zu generieren. Wortbeispiele können aus Texten ausgeschnitten und zu Übungseinheiten zusammengestellt werden. So entstehen Datenbanken für schwierige Lesungen. Sie lassen sich durch Animationen, in denen Schreibbewegungen einzelner Buchstaben in Szene gesetzt werden, sinnvoll ergänzen. Fehlen die technischen Voraussetzungen, das Schreiben einzelner Buchstaben als Programm zu animieren, bieten Videosequenzen eine praktikable Alternative. Mit dem Microsoft Office Werkzeug »OneNote« z.B. lässt sich die Schreibbewegung am Monitor per Mouse-Bewegung darstellen und mit einer Videokamera filmen. Zu Lektionen zusammengefasst, in denen sich Transkriptionsaufgaben zu Wortbeispielen mit Animationen von Schreibbewegungen abwechseln, lassen sich einzelne Buchstaben gezielt und effizient üben. Sich den jeweiligen Duktus zu vergegenwärtigen ist die wirksamste Methode, Buchstaben wie das »d« ( ) von dem »s« ( ) am Wortende sowie das »s« ( ) in der Wortmitte von den Buchstaben »h« ( ) und »f« ( ) zu unterscheiden. Diese Buchstaben sehen sich auf den ersten Blick sehr ähnlich. Macht man ihre gegensätzliche Schreibrichtung deutlich, werden sie leicht unterscheidbar. Studierende sind sich dieser Verwechslungsgefahr oft nicht bewusst und erfahren hier durch das Tafelwerk kaum Unterstützung, da allein der Hinweis auf ähnliche Buchstaben, wie z.B. Süß sie angibt, noch nicht zu ihrer Unterscheidung führt. Hier kommen die unterschiedlichen Lerntypen zum Tragen. Für den haptisch orientierten Lerntyp erschließt sich dieser Zusammenhang nicht allein durch das Betrachten von Bildern. Erst das Beobachten der animierten Schreibbewegung und das anschließende aktive Nachvollziehen (Schreiben) führen zum Verständnis.

Wie erwähnt unterstützt das Tafelwerk nicht das Sinnverstehen. Dieses Lernziel lässt sich hingegen mit digitalen Ressourcen erreichen. Mit ihrer Hilfe können Lesestrategien anschaulich vermittelt werden, die das Erschließen des Sinns fördern, nämlich: Schlagwortsuche und unleserliche Hand. Diesen beiden Strategien kommt im Seminar besondere Bedeutung zu, weil ihre Beherrschung die Studierenden davor bewahrt, sich von schwierigen Lesungen entmutigen zu lassen, sondern sich an die Transkription Stück für Stück heranzutasten. Sie fordern dazu auf, unsichere Lesungen gezielt anzugehen, und setzen geradezu voraus, dass die Lesekompetenz für die vollständige Transkription eines Schriftstückes noch nicht ausreicht. Nicht nur bei der Vermittlung sondern auch bei der Anwendung der Lesestrategien ist die Präsenzveranstaltung wichtig, denn der Leistungsstand kann am besten im Seminar vom Teilnehmer selbst und vom Dozenten überprüft werden.

Durch die Kombination der Lesestrategien und der digitalen Werkzeuge ergeben sich weitere Vorteile, etwa beim Transfer vom Seminarraum in die Klassenräume. So wird der Umgang mit Quellen im Schulunterricht auch dann möglich, wenn keine Zeit zum Erlernen von Handschriften vorhanden ist. Mit Lückentext und individuellem Leseschlüssel lassen sich Quellen so aufbereiten, dass sie innerhalb einer Schulstunde entziffert werden können. Abbildung 2 zeigt einen Ausschnitt aus dem Meisterbuch der Potsdamer Seifensieder und Abbildung 1 den daraus gewonnenen Leseschlüssel für die Buchstaben S bis Z.

Auch die digitalisierten Quellenbeispiele können das Tafelwerk übertreffen. Denn aus den digitalen Materialien heraus lassen sich Erfahrungen im Umgang mit Quellengattungen wie Akten oder Urkunden, die bislang erst im Archiv gemacht werden konnten, schon im Seminar vermitteln. Ein Gefühl für die Charakteristika dieser Quellengattungen stellt sich erst nach Sichtung dutzender Exemplare ein. Die Digitalisierung bietet die Möglichkeit, diese Fülle im Seminar bereit zu halten und somit beispielsweise den Geschäftsgang einer Behörde nachvollziehbar zu machen. So kann auch in puncto Digitalisierung die Quantität in Qualität umschlagen. Genau ein solcher Geschäftsgang ist auf dem Internetauftritt des Geheimen Staatsarchivs Preußischer Kulturbesitz abrufbar. Anhand eines Schriftwechsels des Generaldirektoriums mit den ihm unterstellten Behörden zum Seidenbau werden in der Transkription auch Randnotizen, Paraphen oder Kanzleivermerke aufgeschlüsselt. Auf diese Weise lernt der Nutzer, eine Ausfertigung von einem Konzept zu unterscheiden und kann so Aussagen zum Quellenwert eines Schriftstücks treffen. Neben dem Wissenszuwachs ist zugleich als ein weiteres Lernziel der Kontakt zu Archiven geebnet.

Nicht nur die auf der Lernplattform bereitgestellten Informationen lassen sich durch externe Ressourcen sinnvoll ergänzen, sondern auch die dort eingestellten Übungen. Unter der Vielzahl an Lernprogrammen zur Paläographie dominieren solche zur lateinischen Paläographie des Mittelalters. Dies wird auch an der anschaulich bebilderten Sammelrezension von Georg Vogeler zu historischen Hilfswissenschaften im Internet deutlich (Vogeler). Das Lesetraining von Thomas Frenz befindet sich noch in Vorbereitung und konnte daher hier nicht ausgewertet werden. Die Funktionalität des bereits erwähnten Sütterlin-Lernprogramms SLP 2000 beschränkt sich derzeit leider auf das Anzeigen von Quellenbeispielen, eine Eingabe der Transkription ist nicht möglich. Daher sei an dieser Stelle nur auf zwei Lernprogramme ausführlich verwiesen, an denen sich die Umsetzung zuvor erwähnter Lesestrategien und Lernziele diskutieren lässt: das Projekt der Universität Zürich »Ad fontes« sowie die »Early-modern Paleography« von David Postles.

Ein wesentlicher Vorteil des Lernprogramms »Ad fontes« ist die direkte Eingabekontrolle. Transkriptionsübungen können nicht nur nach den Schwierigkeitsgraden leicht, mittel oder schwer ausgewählt, sondern auch jederzeit unterbrochen und zu einem späteren Zeitpunkt fortgesetzt werden. Dies wird durch die Anmeldung als Benutzer möglich. Beim Transkribieren kann dank der Eingabeprüfung immerhin in Ansätzen nach der Lesestrategie »Lückentext« vorgegangen werden. Das Programm erkennt zwar, wenn ein Wort vergessen wurde, doch kann nicht an einer beliebigen Stelle im Dokument mit der Transkription begonnen werden, wie es die Vermittlung dieser Lesestrategie im Seminar vorsieht. Hingegen findet sich das Lernziel Sinnverstehen innerhalb dieses Programms in mehreren Übungen zur Datierung oder zum Rechnen mit verschiedenen Maßen und Währungen geradezu vorbildlich verwirklicht. Hier wird der Nutzer Schrittweise bis zur kompletten Transkription geleitet. Am Anfang steht das Identifizieren der betreffenden Textstellen. Nach erfolgter Eingabe werden sie im Dokument graphisch hervorgehoben, dann erst geht es mit der Transkription der identifizierten Textstellen weiter. Die Übungen sind sehr anschaulich gestaltet und werden durch Ressourcen und Zwischenergebnisse gut begleitet. Zudem potenziert die Aufteilung dieser Übung in kleine Schritte die Lernmotivation. So ist es einerseits möglich, trotz eines falschen Teilergebnisses noch auf das richtige Endergebnis zu kommen und andererseits wird der Anwender nicht nur einmal für das Lösen einer komplexen Aufgabe belohnt, sondern für jeden richtigen Lösungsschritt. »Ad fontes« sieht derzeit noch kein gezieltes Üben einzelner Buchstaben und Wörter vor, so wie es schon jetzt die »Early-modern Paleography« erlaubt.

Zwar beschäftigt sich dieses Programm auch nicht mit Kurrentschriften, sondern mit Entstehung und Weiterentwicklung der englischen »Secretary hand«, doch hält es eine gelungene Verbindung von Transkriptionsübungen zu einzelnen Worten und Buchstaben mit dem Vollregest bereit. Die Wortbeispiele können unter der Rubrik »Quiz« aufgerufen werden. Hier gibt es nicht nur eine Eingabekontrolle, sondern an diesen Wortbeispielen werden auch Besonderheiten der »Secretary hand« vermittelt, wie etwa ein seitenverkehrtes »e«. Die Wortbeispiele sind den Faksimiles der neun Lektionen entnommen, die das Gros der Transkriptionsübungen dieses Programmes beinhalten.

Der Nachteil dieser Lernprogramme ist jedoch, dass sie eine stehende Internetverbindung erfordern. Zudem begegnet einem hier ein häufiges Charakteristikum digitalisierter Handschriften: die große Datenmenge. So erfordert insbesondere die Nutzung von »Ad fontes« einen schnellen Internetzugang.

5 Perspektiven

Die Nutzung digitaler Ressourcen vereinfacht die paläographische Lehre nicht nur, sie eröffnet auch neue Möglichkeiten, gerade weil die Individualität des Lernens berücksichtigt werden kann. Nicht jeder Dozent ist in der Lage, die in diesem Artikel skizzierten Übungen und Anwendungen zu den vier Lesestrategien selbst zu generieren. Umso wichtiger sind Lernprogramme bei der Verwirklichung der Lernziele. Selbst ein zu Recht viel gelobtes Lernprogramm wie »Ad fontes« schöpft noch nicht das Potential digitaler Ressourcen voll aus. So ließen sich beispielsweise nach dem Vorbild der Lesestrategie »Lückentext« Transkriptionsübungen erstellen, bei denen an beliebiger Stelle im Dokument mit der Übertragung der Handschrift begonnen werden kann. Dies ließe sich auch mit Übungen zum Sinnerschließen verbinden, indem Schlüsselstellen im Text zuerst zu transkribieren sind. Des Weiteren sind auch für die Kurrent- und Kanzleischrift Transkriptionsübungen zu einzelnen Worten und Buchstaben wünschenswert, wie sie bereits für das Programm »Early-modern Paleography« existieren. Die Animation von Schreibbewegungen stellt ein weiteres, bislang ungenutztes Mittel der Veranschaulichung für die Paläographie der Frühen Neuzeit dar. Ebenso fehlen für die Paläographie der Frühen Neuzeit derzeit noch Lernprogramme, die sich als Download oder über CD-ROM auf dem eigenen Rechner installieren lassen.

Ein Desiderat ist weiterhin ein Zeichensätzen zur Kurrent- und Kanzleischrift, ähnlich dem bereits existierenden Sütterlin-Font. So kann der Zugang zu Handschriften vom 16. bis zum 18. Jahrhundert entscheidend vereinfacht werden. Bislang ist noch kein solcher Font im Internet frei verfügbar. Er ließe sich jedoch für kleines Geld über einen Dienst wie YourFonts generieren.

Das kooperative Lernen mit Hilfe der virtuellen Lernumgebung lässt sich ebenfalls noch verbessern. Auf diese Weise können Studierende untereinander ihre persönlichen Lesekompetenzen und ihre Schwächen wahrnehmen. Auf die Gruppenarbeit ist das Design der Lernplattformen noch zu wenig eingestellt.

Des Weiteren kann die paläographische Lehre vom Ausbau von Computerplätzen nur profitieren, da interaktive Online-Lernprogramme den nachhaltigsten Effekt auf Lernverhalten, Lernmotivation und damit auch auf den Lernerfolg versprechen. Der Einsatz digitaler Ressourcen in der paläographischen Lehre fördert mehr als nur die Motivation der Studierenden. Digitale Materialien und Werkzeuge machen die Vermittlung paläographischer Kenntnisse anschaulicher und damit effektiver. Von diesen Vorzügen profitiert aber nicht nur die Paläographie, sondern die universitäre Lehre allgemein. Ein leicht zu beobachtender Lernerfolg, wie er sich beim Transkribieren von

Handschriften einstellt, wirkt sich positiv auf die gesamte Studienleistung aus. Weder Tafelwerk noch Lernprogramme machen die paläographische Lehre überflüssig. Nur durch den Wechsel von Präsenzveranstaltung und E-Learning erhalten die Studierenden eine ausreichende Motivation und Begleitung durch den Dozenten. Gerade diese hier umrissene Fülle an Anwendungen und Potentialen digitaler Ressourcen macht den Wert der Paläographie in der Hochschullehre deutlich: Die Paläographie vermittelt den Studierenden Spaß am Lesen alter Handschriften, indem sie zeitgemäße Medien in das Seminar einbindet.

Bibliographie

Ad fontes. <http://www.adfontes.uzh.ch/1000.php>.

Benutzungshinweise Geheimes Staatsarchiv Preussischer Kulturbesitz. <http://www.gsta.spk-berlin.de/benutzung_3.html>.

Frenz, Thomas. [*Paläographisches Lesetraining*]. <http://www.phil.uni-passau.de/histhw/palaeographie/index.html>.

Geldsetzer, Sabine und Meret Strothmann. »Blende(n)d Lernen in Bochum. Integration von E-Learning in den BA/MA-Studiengang Geschichte an der Ruhr-Universität Bochum.« *Geschichte lehren an der Hochschule. Reformansätze, Methoden, Praxisbeispiele.* Hrsg. Rainer Pöppinghege. Schwalbach: WOCHENSCHAU Verlag, 2007. 181-193.

Grun, Paul Arnold. *Leseschlüssel zu unserer alten Schrift.* Nachdruck der Ausgabe Görlitz 1935. Limburg an der Lahn: Starke, 1984.

»Ich fürchte, wir haben ein Problem.« Erfahrungsberichte von der Universität Potsdam. Antworten auf einen Aufruf der Studentischen Vertreter im Senat der Universität Potsdam. 2009. <http://www.pep.uni-potsdam.de/media/publikationen/up_aufruf-erfahrungsberichte_mwfk_090129.pdf>.

Mankel, Mirco. *Lernstrategien und E-Learning. Eine empirische Untersuchung.* Hamburg: Kovač, 2008.

[*Meisterbuch Seifensiedergewerk*] Stadtarchiv Potsdam 1-12/169, fol. 17.

Postles, David. [*Early-modern Paleography*]. <http://www.paleo.anglo-norman.org/palindex.html>.

SLP 2000 [Sütterlin Zeichensatz]. <http://www.phil.uni-sb.de/projekte/suetterlin>.

Süß, Harald. *Deutsche Schreibschrift lesen und schreiben lernen.* Augsburg: Weltbild, 2004.

Vogeler, Georg. »Historische Hilfswissenschaften.« *eLearning Mediävistik.* Hrsg. Hiram Kümper (im Erscheinen).

YourFonts. [Online-Font-Generator]. <http://www.yourfonts.com>.

Digistylus — An Online Information System for Palaeography Teaching and Research

Antonio Cartelli, Marco Palma

Abstract

This paper starts by describing the experiences the authors recently had with online information systems for teaching and research in palaeography. The study also considers the differences in the students' access to the site "Teaching Materials for Latin Palaeography" when they attended the palaeography courses, as it was usually used in the lectures by one of the authors. With the increase in the quantity of plates (reproducing pages or parts of them from medieval manuscripts) and texts (concerning the analysis of the writing styles, the cataloguing, the history of manuscripts, the codicology and other important topics in the palaeography's scientific debate), it became clear that there was a difference in the way students approached those materials: when students first used the systems in the academic year 2001–2002, they read all the documents and used all the plates; more recently, with the quantity of materials on the site considerably increased, the students wait for the professor's suggestions and evidence uncertainties and difficulties when autonomously looking for a document or a plate. As a consequence, the online information system Digistylus has been planned and is going to be created for the management of the data in the site "Teaching Materials". The main consequence of the above observations has been the detection of a new knowledge construction paradigm and the development of new research procedures in palaeography.

Zusammenfassung

Der Artikel beschreibt die Erfahrungen der Autoren, die sie kürzlich mit dem Online Informationssystem für Lehre und Forschung zur Paläographie gesammelt haben. Die Studie betrachtet insbesondere die Unterschiede, die sich beim Zugang der Studenten zur Website »Teaching Materials for Latin Palaeography« während ihrer Paläographiekurse zeigten. Mit steigender Anzahl der angebotenen Bilddigitalisate mittelalterlicher Handschriften und von Texten zu Schreibstilanalyse, Katalogisierung, Manuskriptgeschichte, Kodikologie und anderen wichtigen Bereichen paläographischer Forschung wurde der Unterschied deutlich: die ersten Studenten im Studienjahr 2001–2002 lasen alle Dokumente und benutzen alle Bilddigitalisate. Im letzten Studienjahr, in dem die Anzahl des verfügbaren Materials deutlich angestiegen ist, warteten die Studenten auf

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 123–134.

die Vorschläge des Dozenten und zeigten Unsicherheiten und Schwierigkeiten, wenn sie eigenständig nach Dokumenten oder Bilddigitalisaten suchten. Infolge dieser Beobachtungen ist das Online-Informationssystem Digistylus für das Datenmanagement der Website »Teaching Materials« geplant. Das Hauptergebnis der obigen Beobachtungen ist die Entdeckung von neuen Wissenskonstruktionsparadigmen und das Finden von neuen Forschungsvorgängen in der Paläographie. Als Hauptergebnis der Studie kann die Beobachtung eines neuen Paradigmas zur Wissenskonstruktion sowie neuer Herangehensweisen im Bereich paläographischer Forschung genannt werden.

1 Introduction

Since 2001 the authors have worked on the construction of special web sites (mostly online information systems), used both for research and teaching. The spread of the Internet and the easy use of the web for retrieving information and putting new data in a given database were the main reasons for the construction of the online information systems. The sites described below have been used to manage bibliographical data on medieval manuscripts; they also implemented the processes usually adopted by researchers for the collection of bibliographical data.

Women and written culture in the Middle Ages (Cartelli et al.), contains the names of the women who wrote manuscripts in the Middle Ages, and the manuscripts they wrote; when suitable images are available, people accessing the site can also see the women's handwriting styles. The database can be accessed by the persons who planned the system and by those authorized to input the bibliographical data, the images, the bibliographies etc. and can be queried by everyone.

The *Open Catalogue of Manuscripts of the Malatestiana Library* (Cartelli and Palma; Cartelli et al.), is the most complex system among the ones created until now (it derives from the more general idea of the *Open Catalogue* and has been created by the staff of the Malatestiana Library). Among other things, it lets people know the history of the library and provides open access to all available pages in the manuscripts[1].

BMB on line (Cartelli and Palma) is a pure bibliographical information system; it manages the quotations of Beneventan manuscripts. People engaged with the collection of the quotations of those manuscripts are grouped into three categories: contributors, who can access web forms by writing, modifying and deleting bibliographical data; scientific administrators, who can manage all the data that the contributors are charged with and write, modify, and authorize bibliographical materials (i.e. this last operation can be done only once, because authorized records cannot be reviewed by

[1] See the paper of Cartelli et al. "Il catalogo aperto dei manoscritti Malatestiani" in this volume, to get an idea of the different sections and autonomous information systems it is made of.

contributors and scientific administrators but can be retrieved by general users); the system administrator, who is allowed to do all operations, including the modification or deletion of authorized bibliographical records.

General users can access authorized materials on the site according to different query pages: by author, by manuscript, by contributor, and by one or more words or part of them in the title, the location, or the bibliographical abstract of a given publication.

It has to be noted that there are different elements also implemented within the system:

- *a closed communication subsystem*: it can be accessed only by people working on the information system (contributors, scientific administrators and system administrator) allowing the quick exchange of messages and texts;
- *special functions*: available only to the system administrator, for the production of printed versions of the collected data (to be used to create a printed publication concerning the bibliographies collected yearly).

Until now all the web sites and the information systems mentioned above have been used for research and teaching in palaeography and the following effects have been detected (Cartelli 2006a, 2007):

1. Every group of persons working on a given information system (students, professors, researchers etc.) showed the features of a community of practice as described by Wenger (2004): people identifying themselves in the community they belong to, people having a common and shared commitment, people sharing special signs, symbols and strategies (i.e., the repertoire of the knowledge instruments in the community).
2. The sites have been good examples of constructivist learning environments and have helped the students to develop cognitive apprenticeship strategies (very useful for the improvement of their learning and performance).
3. The features of communities of learners (CoLs) and fostered communities of learners (FCL) were detected in the classes working on the systems described; otherwise stated, the online information systems, while supporting and extending traditional learning strategies, induced the creation of special communities, which is never detected in traditional palaeography courses.
4. New skills emerged in the students involved in the experiences described above, such as the ability for team work, the management of complex tasks and the raising of the individual's skills within the community.
5. New transversal competences were detected: better computing skills of the students attending traditional computing literacy courses and meta-cognitive / cognitive apprenticeship strategies.

It should be noted that the greater benefits for the students (points 2–5 above), compared to researchers (point 1), is mostly due to the desire of involving them in the use of new ways of learning, so that the sites were used for teaching very early. Comparing the performance of these students with those of traditional students and qualitatively analysing the simple questionnaires and the interviews at the end of the lectures, a similarity of the results reported above with those of the North American researchers on communities of learners has become clear.

Similar procedures could not be adopted with the scholars and researchers who worked on the systems, because most of them collaborated with students for the creation of special documents or images only occasionally. In these last cases the simple observation of the individuals' behaviours and qualitative enquiries were used to deduce the presence of the features of communities of practice in the groups of persons working on the information systems.

2 Information Retrieval and the Digistylus System

The online information systems described in the first section of this paper have both been used for research and teaching, but the need for helping the students to easily access teaching materials led to the creation of the static web site "Teaching Materials for Latin Palaeography". It aimed at making the materials for the understanding of the ancient writing styles available to the students, and to prepare them for their final examination in Latin palaeography (Cartelli and Palma 2005).

The site, which continuously evolves, comprises three sections:

- The first section contains the plates, reproducing folios of the ancient manuscripts (texts written in different medieval scripts); together with the images there are the transcriptions (i.e. digital full texts where symbols, special signs and abbreviations are clearly written). All the documents are organized as a tree structure, based on the writing style adopted in the different plates.
- The second section contains full or partial documents reproducing papers, presentations and articles (from conferences, catalogues and books), on different topics such as book archaeology, scripts, cataloguing, history of palaeography etc.
- The third section is used for work in progress, it hosts special documents, usually simple archives created with office automation programs (like MS Excel or MS Access); they are managed by the professor and students (and can be downloaded from the site).

The web site began in 2001 and helped students to develop skills for reading and understanding ancient handwriting styles, learning the history and the evolution of European national languages (especially Italian) and for dealing with the processes, the strategies and the policies for the preservation of ancient manuscripts.

A relevant change in the way the students accessed the materials on the site was detected during the last years by observing students' behaviours and by asking them (when attending final examinations) about the pages and materials they visited and the time they spent working on them.

The main unexpected conclusion was that the more the materials, the more the difficulties the students had in autonomously managing the study materials. The following two opposite behaviours emerged from all students' answers: at the beginning (i.e. when only a few documents were available), the students read all the texts and autonomously transcribed almost all the plates (then compared the text they produced with the professor's solution); later (during the last course), when 86 documents and 281 plates with their transcriptions were available, the students mostly read only the texts the professor suggested in his lectures and limited themselves to the analysis of the plates they discussed in class.

What are the reasons for the last behaviour? Students say that they have trouble finding the "right documents" to study or to analyse when they autonomously browse the site. Very often, in fact, they have to read more than one document before finding the right information or before understanding what document to search for and, sometimes, this time-consuming job prevents them reaching their goal.

The points below can be useful to outline the reasons for the changes in students' behaviour:

- The increase in the quantity of materials in the site,
- The overestimation of the students' knowledge and skills when they are requested to find information,
- The generational differences and the approach which younger students have to technology.

It is not hazardous to include the above points in a wider hypothesis by which we are facing a more general problem concerning the search of information on the web.

Perhaps the solution can be found by asking the following questions: what difficulties do people have while searching materials on the web? Do they succeed in finding the right data on the web? How can the students be helped in searching for the information they need and to build new and meaningful conclusions?

The last question does not completely include all the other questions but it is the most comprehensive one, so that we'll concentrate on it in the following section. In this regard the Digistylus information system, which hopes to address the question, will be discussed.

2.1 The Information System Digistylus

The planning and execution of an information system which could help students to easily access the plates and the documents available on the site of the "Teaching Materials for Latin Palaeography" looked like the best solution to the problems that students had in retrieving documents.

Many different considerations guided the creation of the information system and some of them are reported here: students must be the creators of the information on the site (they must organize and input in the system all the data concerned with the documents in the site); the information in the database must be available to everyone who may be interested in it (by means of the web); any information the students put in the system must be approved by one or more scientific coordinators before being available on the web; special indices must be implemented in the system to let people measure the difficulty in the transcription of the plates; a closed forum within the information system is needed, which lets students communicate among themselves and with the professor; the evaluation of the students' work and the final score they obtain at the final examination must include different elements: the evaluation of the work he/she made up, the evaluation of the support he/she gave to colleagues, the evaluation of the accessibility and usability of the information retrieved by external readers (i.e. general users), the evaluation of criticism he/she gave to the system and its functions.

It can be easily recognized that students are involved in the project at different levels:

- Individually: by critically studying and assimilating the basic topics of the discipline, by applying those ideas to the materials on the site and by writing the records in the database (this job is made easier by the presence of supporting materials and the use of communication subsystems).
- At a community level: by adopting various strategies: a) the legitimate peripheral participation (LPP) suggested by Lave and Wenger (1991), helping the management of the community while including the weakest subjects, b) the implementation of practices with the Information Communication Technology (ICT), proposed by Cartelli (2008), letting the system implement the processes people had to conform to, and governing the management of the information acquisition, storing and validation, c) team competency learning, suggested by Jewels and Albon (2006), inducing the professor to act as a coach and assign to every student the best role with respect to his/her basic knowledge and skills.
- Socially: by considering the usefulness of the information the students produce for their professor and their community, but especially to the people not necessarily expert in Latin palaeography or in any other discipline concerning the study of ancient manuscripts.

The structure of the information system based on the ideas mentioned above is outlined in Figure 1, where a snapshot of the data structure and the flow of information is drawn.

Before any other consideration the following remarks should be made: the creation of the Digistylus information system leaves the former site "Teaching Materials for Latin Palaeography", with all the documents and the plates within it, unchanged; people who

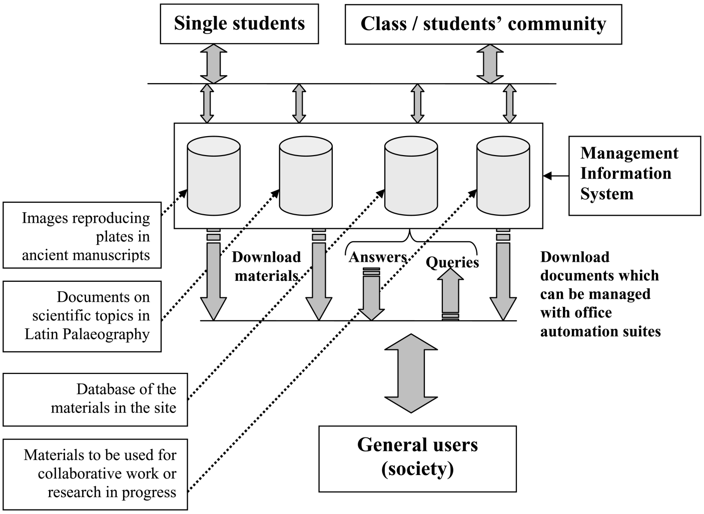

Single students
Class / students' community
Management Information System
Images reproducing plates in ancient manuscripts
Documents on scientific topics in Latin Palaeography
Database of the materials in the site
Materials to be used for collaborative work or research in progress
Download materials
Answers
Queries
Download documents which can be managed with office automation suites
General users (society)

Figure 1. The Digistylus information system built around the site "Teaching Materials for Latin Palaeography".

like to access those materials in a more traditional way can do so by using the links in the web pages and browsing the site.

The Digistylus information system has been created with Open Source software, which is used for the web server (Apache), the Relational Data Base Management System (PostgreSQL) and the creation of special web pages interfacing the database (PHP).

Digistylus' structure is based on one database, a sophisticated user rights management, a query system and a specific data flow. In what follows, the above elements are explained in greater detail:

1. The relational database underlying the site is made of tables containing the following data:

    - Contributors' (students) and scientific administrators' personal identification data.
    - Shelfmarks of the manuscripts containing the plates reproduced in the site.
    - The bibliography of the manuscripts and of the medieval documents in the site.
    - The graphic style of the plates and all the data (as far as they are available) which can be used for a better description of those plates.
    - The links to the web pages with the reproduction of any plate and its transcription.
    - The keywords letting people access the transcription of a given plate.

- The parameters for the calculation of the difficulty in making the transcription of a given plate.
- The bibliographic records of the documents in the site, with the links to the corresponding documents.
- A communication subsystem which allows people working on the system to communicate, manage bibliographical records and input new data in the Digistylus database.

2. The users accessing the database have different roles and permissions: the users with the lowest permission on the data are the ones who can only query the system; they can see the plates, the transcriptions, the list of the bibliographies and any other information on the site, but they cannot insert or modify information in the database. At the next level are the contributors (students) who can access a special web area (by means of their ID and password) with a menu of the allowed operations (i.e. they can manage the records on the plates and their transcriptions, the bibliographic cards and the electronic blackboard). At last the scientific administrator/s can manage all the data in the database and write, modify and authorize the bibliography. At the top of the access pyramid is the system administrator who can conduct all the operations allowed to the scientific administrator/s and can access the verified bibliography to modify or to delete it.
3. When the record on a given plate is written, the transcription is prepared and the bibliography is compiled, the scientific administrator can verify and authorize it. All these data can then be queried by a general user. People interested in the information have different query pages:
    - The first queries the author of one or more texts in the manuscripts (the list of the links to the web pages with the plates and their transcriptions completes the system answer);
    - The second lets the user select the author of one or more catalogues and gives back the list of the catalogues used for the reproduction of the plates (like in the former case the list of the links to the web pages with the plates and their transcriptions completes the system answer);
    - The third lets people search for key words (or parts of them) in the transcription of a plate and gives back the list of the plates containing them;
    - The fourth and last page lets the user type in a query the name of the author of a text, the topic, words in the title or in the text and search for any document in the web site which respects these constraints.
4. When the system starts, the data base is empty and the system administrator has to input the data for at least a scientific administrator; once a scientific administrator is enabled, he/she can input the data for one or more contributors and give them access to the system; he can also input the records on the plates and their transcrip-

tions and the bibliographic data. Then the contributor can compile the bibliography in the database. Finally the bibliography is analysed, revised and verified by the administrator/s and can be read and queried by a general user.

It is desirable that the Digistylus system, when completed and ready to use, will contribute to change teaching and research methods in palaeography.

The reasons for the prospected changes are described in the next section.

3 New Paradigms for Knowledge Construction and Palaeography Research

In the introduction the effect the information systems had on students' performance and skills was described. The idea that online information systems influenced the creation of learning communities led to the hypothesis that new technologies changed or at least introduced new approaches to the construction of knowledge.

By following Ong's (2002) and Olson's (1991) ideas of a connection between technology, literacy and new orality, new possibilities for human communication and knowledge construction could in fact depend on the use of the ICT.

The following remarks describe the features of the different levels of influence of the ICT on knowledge construction by looking at two different points of view, the personal development of knowledge and a more theoretical one.

In the first case, the subject's point of view is considered. From this perspective, people build their knowledge in three different ways. The first is the autonomous interaction with phenomena, whether they are real or virtual (mostly constructive). The second is the social interaction with other individuals in a community, where mediation, interpersonal contacts, informal knowledge sharing and support from peers have a more relevant role (and ICT are important in helping subjects create communities or induce communities). The third implies the active participation in the society they are immersed in (with respect to community, emulation of behaviours as well as codified and socially accepted rules may modify pre-existing learning strategies or determine new ones). As a conclusion, subjects' knowledge is made of three components: the individual, the community and the social ones, with their own contents, learning strategies and possible communication channels (Cartelli, 2006b). Figure 2 gives a snapshot of the tri-partition of this viewpoint.

The second viewpoint is concerned with the analysis of knowledge by itself. Knowledge is now seen as a theoretical construction, or an artifact of mankind. In this case, like the former one, at least three kinds of knowledge can be recognized. The first is individual knowledge, built by subjects who construct their knowledge while interacting with the environment they are immersed in (natural or virtual, populated by other subjects or not etc.). The second is community knowledge, belonging to com-

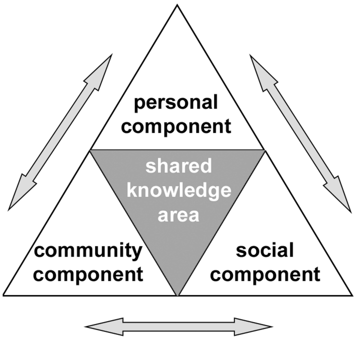


Figure 2. Knowledge components and their interactions.

munities as autonomous entities (by using Wenger's words, it is the knowledge letting communities identify themselves in an autonomous social environment, where people have common aims and motivations and share a repertoire of instruments, which are made of signs, symbols, processes and strategies). The third is society knowledge (often called scientific knowledge), which is well codified, evaluated and approved by a relevant number of individuals and communities (it can probably be identified with the scientific knowledge or with its paradigms).

Figure 2 illustrates the described situation.

The main results one can deduce from the considerations above are:

- Knowledge construction is the result of the influence of all three components.
- Planning and carrying out an information system for the management of information must consider all the knowledge components described until now.
- Using information systems for the implementation of the practices shared by a group of specialists or by subjects working together can be considered a new pedagogical paradigm; it forces other people (students, general users etc.) to create a new community or enter into the already existing community and displaces the problem of "information research" with that of "information creation".

In the introduction the consequence for the use of the new teaching paradigm on the students has already been discussed, now a question sounds interesting: how much is palaeography research influenced by the knowledge construction tri-partition proposed above?

At least two very important effects can be hypothesized:

- Before the Internet and the use of information systems for data management, scholars of Latin palaeography mostly interacted with knowledge they used for studying ancient manuscripts alone (social knowledge); now new relationships arise, at least among people in the community they belong to (i.e. when special information systems are adopted), and community knowledge becomes an important element to guide research in the discipline topics,

- Whatever the role knowledge plays in the development of community processes for study and research, the subjects in the community working surrounding an information system tend to specialize and to construct research teams in order to avoid conflicts and have the best results.

Another relevant effect of the knowledge tri-partition hypothesis is the opening towards the society of the researcher work and its transparency. It has, in fact, a double influence on the management of activities. Firstly, those who are interested in the information in the site can give a feedback to system administrators; new information, new hypotheses and applications will be the starting point for new research. Secondly, the control of process management is possible by means of the results available; everyone can verify the evolution of a research project and the reaching of the pre-defined goals.

It is too early to say whether the above ideas will have impact on more traditional research methods and especially on qualitative and quantitative methods, and future investigations are needed to analyse the results of the use of information systems in palaeography.

Bibliography

Cartelli, Antonio. "TIC e didattica: due esperienze a confronto." *E-learning. Formazione, modelli, proposte.* Ed. Paolo Crispiani and Pier G. Rossi. Roma: Armando, 2006a. 145-154.

Cartelli, Antonio. "Semantics, Ontologies and Information Systems in Education: Concerns and Proposals." *Journal of Issues in Informing Science and Information Technology: The Information Universe* 3 (2006b): 113.

Cartelli, Antonio. "From Socio-Technical Approach To Open Education: MIS and ICT for the Definition of New Teaching Paradigms." *Proceedings of ECEL 2007 International Conference.* Ed. Dan Remenyi. Reading: Academic Conferences Limited, 2007. 97-106.

Cartelli, Antonio. "The Implementation of Practices with ICT as a New Teaching-Learning Paradigm." *Encyclopedia of Information Communication Technology.* Ed. Antonio Cartelli and Marco Palma. Hershey (PA): Information Science Reference, 2008. 413-418.

Cartelli, Antonio et al. "The Open Catalogue of Manuscripts of the Malatestiana Library." *Encyclopedia of Information Communication Technology.* Ed. Antonio

Cartelli and Marco Palma. Hershey (PA): Information Science Reference, 2008. 656-661.

Cartelli, Antonio, Luisa Miglio, and Marco Palma. "New Technologies and New Paradigms in Historical Research." *Informing Science, Special Issue "Widening the Focus"* 4 2 (2001): 61. <http://inform.nu/Articles/Vol4/v4n2p061-066.pdf>

Cartelli, Antonio and Marco Palma. "Towards the Project of an Open Catalogue of Manuscripts." *Proceedings of IS 2002 Informing Science + IT Education Conference.* Ed. Eli Cohen and Elizabeth Boyd. Santa Rosa (CA): ISI, 2002. 217-224.

Cartelli, Antonio and Marco Palma. "The Open Catalogue of Manuscripts Between Palaeographic Research and Didactic Application." *Proceedings of the IRMA 2003 Conference "Information Technology & Organization: Trends, Issues, Challenges and Solutions".* Ed. Mehdi Khosrow-Pour. Hershey (PA): Idea Group Publishing, 2003. 51-54.

Cartelli, Antonio and Marco Palma. "BMB on line: An Information System for Palaeographic and Didactic Research." *Proceedings of the IRMA 2004 Conference "Innovation through Information Technology".* Ed. Mehdi Khosrow-Pour. Hershey (PA): Idea Group Publishing, 2004. 45-47.

Cartelli, Antonio and Marco Palma. "Computer and Information Systems in Latin Palaeography between Research and Didactic Application." *Technology Literacy Applications in Learning Environments.* Ed. David Carbonara. Hershey (PA): IGI Global, 2005. 288-298.

Jewels, Tony J. and Rozz Albon. "Teaching Team Competences." *Teaching in the Knowledge Society: New Skills and Instruments for Teachers.* Ed. Antonio Cartelli. Hershey (PA): Information Science Publishing, 2006. 174-186.

Lave, Jean and Etienne Wenger. *Situated Learning. Legitimate Peripheral Participation.* Cambridge (MA): Cambridge University Press, 1991.

Olson, David R. and Nancy Torrence. *Literacy and Orality.* Cambridge (MA): Cambridge University Press, 1991.

Ong, Walter J. *Orality and Literacy: the Technologizing of the Word.* New York (NY): Routledge, 2002.

Wenger, Etienne. *Communities of Practice: A Brief Introduction*, 2004. <http://www.ewenger.com/theory/index.htm>

Innovations in Analyzing Manuscript Images and Using them in Digital Scholarly Publications

Bernard J. Muir

Abstract

Evellum began developing software for the digital analysis and presentation of medieval manuscripts nearly fifteen year ago, when there were very few design and delivery options available to programmers. In the early years, it was not apparent how it would be best to deliver such products nor exactly how they would function and be used, and the question of longevity plagued us. Today there is the TEI to help standardize the mark-up of text and to offer a greater guarantee of longevity than was previously possible, and internet browsers are capable of facilitating the delivery of programmes that integrate text, image and video. Two products designed by Evellum are described here, with comments on the pedagogical issues that have helped determine their shape.

Zusammenfassung

Vor fast fünfzehn Jahren begann Evellum mit der Entwicklung von Software für die digitale Analyse und Präsentation mittelalterlicher Handschriften. Zu jener Zeit verfügten die Entwickler noch über wenige Alternativen bezüglich des Designs und der Verbreitung von Software. Damals waren die Vertriebswege für Softwareprodukte im Wissenschaftsbereich noch nicht so etabliert wie heute, Nutzeranforderungen nicht genügend erforscht und die Funktionsfähigkeit der Produkte über wechselnde Generationen von Hard- und Software nicht gesichert. Heute hilft die TEI, Text-Markup zu standardisieren und eine höhere Lebensdauer der Produkte zu garantieren. Die inzwischen allgegenwärtige Internet-Technologie erlaubt es uns, Programme auszuliefern, die Text, Bild und Video integrieren. In diesem Beitrag werden zwei der von Evellum entworfenen Produkte vorgestellt und die didaktische Fragestellung, die bei ihrer Ausgestaltung half, kommentiert.

Nearly twenty years ago I was appointed the inaugural Associate Dean for Information Technology in the Faculty of Arts at The University of Melbourne. I was then a budding medievalist who knew how to format documents on a personal computer; this made me the local IT expert and was most likely the reason why I had been appointed, though I do not think that the acronym 'IT' was used back then! Times have changed. Today I

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 135–144.

am the Professor of Medieval Studies and for the past fifteen years have run a digital software publishing business, Evellum; as the name suggests, the business focuses on electronic vellum or parchment, that is, the interpretation and processing of ancient and medieval manuscripts for delivery in digital format. I now produce two series of digital publications, the Bodleian Digital Texts series, with Oxford University, and my own Evellum Scriptorium Series, which I produce here in Melbourne.

Younger enthusiasts may not be familiar with the Dark Ages of IT and the challenges that confronted those who wanted to somehow use multimedia in their publications and teaching materials. We decided that Apple computers were the way to go and at that time Hypercard was about all that was available to work with; but it was amazing what people were able to achieve in such a restrictive situation. A couple of years later, Oracle Media Objects (OMO) appeared on the market and had a number of more advanced features to offer than Hypercard, so we migrated. It was at this time we realized that the software was primitive and that it would obviously develop in an incrementally fast manner over the next decade, so we made an important decision, that from this point onwards we would keep the media and the program separate from each other so that our work could be moved to another platform or program relatively easily. In this way we foreshadowed the agenda behind the Text Encoding Initiative. Though there were and still are issues relating to the replacement of superseded mark-up instructions, most of these can be resolved by programming, thus reducing the amount of tedious and repetitive key-stroking.

Advances continued to be made in the IT world and at a brainstorming meeting one day a couple of years after moving to the OMO software, a voice from the back of the room said he thought that the future for all this stuff was the internet. What was that? The net was the new boy on the block and the kind of implementation that was being suggested had not even been contemplated yet. We were all still preoccupied with the novelty of electronic mail and fascinated with how it had caused the world to shrink. Afterwards, we moved to Netscape, which we used as the front end for our first CD ROM. Then came Internet Explorer, which offered more features, and so we moved to it and used it for the next two projects. As is well known, the always strained relationship between Microsoft and Apple eventually ruptured, which had the unfortunate effect of making the products we had developed using IE useless for Mac users unless they kept an out-of-date copy of IE 5.2 on their computer specifically for our products. In the event, we have redone a couple of the earlier products and they now work on two or three very popular browsers such as Firefox and Safari.

But as we all know, in recent years the Text Encoding Initiative has come to the fore and seems now to represent our best chance of producing marked-up text that will be readable for many decades to come in spite of radical changes in delivery methods. For two of our current projects we are now using XML mark-up with the Oxygen software and are able to reproduce the many features which have made our DVDs innovative and an embodiment of cutting-edge technology in the field of manuscript studies. To date, we have produced several digital facsimile editions (see the Bibliography), two of which are the focus of the rest of this paper: "Ductus", a program for teaching the rudiments of Latin palaeography and codicology, and "The Making of a Medieval Manuscript", a documentary film which records the actual making of a manuscript in true medieval fashion from the preparation of the raw materials to the finished book.

Back in the late 80s and early 90s a number of academics, availing themselves of the desktop computer and its potential as they saw it, undertook to produce interactive programs and editions for use in teaching, and a few of these had notable success at the time. Pat Conner's (UWV) *Beowulf* Workstation, developed for a Mac, was to my mind the most successful of these, but I remember a few others such as the Thomas à Kempis *Imitatio Christi* fondly, with its tinkling waters in the courtyard and birds cheeping outside the window. Through their multi-dimensional structure and the use of sound these and a few other such works began to instil in us an awareness of what we were doing as 'multimedia' rather than just basic animation. An unforeseen problem for such academics was the rate at which development in the field would escalate, at the same time becoming more complex. Academics who tried to be a 'Renaissance man' in the Digital Age soon found that they became mired down in the technicalities, which consumed the precious time that they should have been putting into content.

Teamwork, the allocation of tasks to individuals with specialist qualifications to do just one aspect of a project, so common in science and medicine, was foreign to the average Humanities and Social Science academic. This was the second major realization that dawned upon us at an early stage, and I feel that it is what has given us the edge over our colleagues during the past decade. It has led to increased productivity, more publications, and a greater number of research grants in a highly competitive national environment. The research and development group at Evellum consists of programmers and designers, research assistants, concept developers, a project manager, and myself, overseeing the whole of each project; I write the grant applications and correspondence, negotiate contracts, and travel the world spreading the good news at conferences attended by my peers. I also write the final version of the content of each new DVD. To this team can be added a small group of artists and technicians with specialist skills who contribute to individual projects as required. Interestingly, most of these people offer their services free-of-charge—for them it is satisfying and exciting just to be part of such an enterprise.

But there is no use in producing products, no matter how good they may be, if people do not hear about them—the best teaching tool in the world lies idle and ineffective until someone uses it in a classroom or lecture theatre. We have never spent much on advertising; rather we have been patient and allowed our reputation to speak for us. Now people come to us via our website, finding us by using a search engine, perhaps the most invaluable IT tool developed in recent years. We now have momentum and the power to influence future developments in our field, but it took over a decade to get here. Each new product developed advertises on its DVD insert both existing products and those under development, and scholars with specialist skills are now being invited to propose new titles in the Evellum Scriptorium Series; there are now plans for DVDs on 'Inside a Medieval Scriptorium', 'The Vikings and their Heritage' and 'Medieval Music'.

This past year has been interesting in that two of our research assistants have been based overseas, still contributing to our projects while completing work on their own theses; today, of course, this is commonplace, but it is another example of how IT has shrunk the world. As this discussion moves towards issues of design, functionality and didactics, it should be noted that the development of such DVDs has provided a remedy for a dire situation in timely fashion. As is well-known, small specialist subjects (sometimes referred to disparagingly as 'boutique') are today at risk in universities everywhere as they begin to experience budgetary exigencies, tightening of the belt after a decade of rapid growth as a huge influx of mostly Asian students poured vast sums of money into the coffers of Western universities. Recently this trend reversed and the good times came to an end abruptly. Manuscript studies is naturally one of the areas under the microscope where previously it enjoyed an untroubled life. Increasingly, these universities are turning to software developers such as us for solutions to their crises. Rather than employ a senior academic with specialist qualifications to teach these small classes of advance students, it is more attractive to purchase software which can be used by a less-qualified or fractional appointment to teach the subjects. Moreover, I have for a number of years taught palaeography and codicology to students all over the world who do not have access to these subjects where they are, whether in a university, a small town in the American mid-West, or the Mojave Desert (really!). At one stage I ran a postgraduate subject at the University of Calgary for a complete semester without having actually been there.

The remainder of this paper is a description—*cum*—reflection on how two projects were conceived, developed, delivered and received, the final stage being the most important both pedagogically (for users) and psychologically (for us, giving us confidence as we forge onwards). Student satisfaction is a most important consideration for us for two reasons: we obviously want the user to have had a stimulating and challenging intellectual outcome from our programs, but these days government funding is usually linked to outcomes recorded in student surveys.

Project 1: Ductus: Handwriting and Bookmaking in the Middle Ages

Ductus was originally designed to meet a perceived need, to enable people who did not have access to expert instruction, wherever they might be, to learn how to read ancient and medieval Latin handwriting during the period 100-1500 CE. At the time the idea was revolutionary in the field; today there are many websites that offer an introduction to palaeography, but none of them seems to be as comprehensive and to have the same sort of resources available as are found on the Ductus DVD. Indeed, a cursory review of those sites and courses being taught around the world reveals that many have modelled their work on our program.

In deciding what should be on such a CD / DVD, the essential items were thought to be: 1) a set of very high resolution images; 2) detailed analyses of each script; 3) annotated transcriptions of each facsimile; 4) glossaries of various sorts (terms, types of manuscripts, library codes); 5) video clips showing how some of these scripts were written, so that students could see the actual 'ductus' of each letter (hence the title of the program); a semester-long course; various 'support documents' and forms to assist users when completing their weekly assignment; and an electronic / virtual library. This last item was ahead of its time, once again. We were proposing to scan essential articles and create a virtual library on the disk so that people working in small institutions without a specialist collection of books on manuscripts or else in remote locations would have access to requisite reference materials. No such thing had at that time been contemplated or at least implemented by publishers so they did not know how to respond to our request. They just said no, either because they were suspicious of what we wanted to do or because they had no administrative system in place to charge us (not to mention they would not have known how to calculate the fee). Stymied at first, we ultimately decided that they indeed had the right to charge us for the reproduction of an entire article, but they would have no recourse if we merely 'summarised' the papers, which is what we did. Problem solved.

At the foundation of all our work is a firm belief that in order to produce the best results you have to begin with media of the highest quality. At that time we used 70 MB scans of each manuscript; today, it is 100 MB plus, which provides us with a library of 300 dpi archival materials. To give you an idea of what the implications were for setting such high standards, let me put this into perspective. When I first began using 100 MB scans, for the *Exeter Anthology* DVD in the mid 90s, I had a quite new Mac, with upgrades: when I clicked a file to open it I could go for a coffee because it regularly took about twenty-five minutes to open one image. And I had 250 images to review. Just to ascertain that none of the scans was corrupt took a few months (bearing in mind that this kind of work is done evenings and weekends when routine academic work is finished). For one of my current projects I have 750 100 MB plus images to deal with, but this all now seems merely routine. The top folder of this project contains 80 GB of data; this is backed up on a regular basis on four different external hard drives. The scale of difference between then and now is staggering.

The master image files were then adjusted as required and from them five different-sized sets of images were generated for various uses, being from 'thumbnail' to 'huge'. One set of these was made specifically for mapping, which is perhaps the most tedious aspect of this kind of work, but it is also what gives 'life' or dynamism to our projects:

the more mapping of hotspots, the more the data can be manipulated. Approximately 7,000 hotspots had to be mapped for the Ductus DVD, many more for the Exeter DVD, which contains three times as many images. Anyone involved in the preparation of digital projects will have noticed that one of the major differences between analogue and digital publishing is that with the latter you complete the 'proofreading' stage before the disk goes to production, not afterwards in a 'proof stage'. Once you have typed in a string of data and moved on, it is unlikely that you will ever revisit it. This is because there are just too many links ever to re-check properly (it is in the nature of the beast), so our Golden Rule is as soon as you have finished typing the data for a link, it has to be carefully proofed before moving on. I issue anyone whom I have chosen to work for me with a set of Guidelines: the first dot-point says, "As soon as you blink, stop working on my projects." Sounds blunt, but it is the most important thing that they can do for me; once corrupt data is entered, you can never be certain that it will be discovered until one day after the disk has been released a user gets the fatal 'File not found' message.

On the original Ductus CD, the 14-session course is called 'Course A', which implies 'B, C…', but they are not to be found. Our original idea, and it is still valid in principle, was that as much data as possible, no matter how basic it might seem, should be entered for each facsimile and its commentary and analysis. This is because at a subsequent stage the data can be sorted and configured in different ways for different user-groups. These groups might either be at different levels, from rank amateur to expert, or have different expectation of needs—for example, someone may wish to extract all the files associated with a particular script, period or region, and configure that into a course of their own. It was originally thought that we might do that ourselves (hence the 'A'), but subsequently seemed like the obvious thing that instructors might want to do for themselves; and this has proven to be the case. But we had foreseen it, which is the point in a discussion of creating a concept for a program.

We thought it essential that there be some sort of video clips to demonstrate the 'ductus' of each letter (the order in which its constituent strokes are made), as I have already mentioned. In our first attempt at this in the mid 90s, before making video clips had become commonplace, we decided to take individual letters of a script and capture their strokes in separate files, so that an 'e' might have 3 basic strokes. These would then be played as a movie and the letter would 'write or create itself' on the screen. Obviously, if you subdivided each of these basic strokes into four partial captures, then when played back the creation of the letter would be less 'jerky' or more flowing and thus more pleasing aesthetically.

We still have those early, experimental movies, but soon afterwards the technology for capturing video and formatting and compressing it was becoming more common and easier to use, and so we moved on. Subsequently, we filmed a calligrapher writing out some different scripts in real time. The beauty of this is that the calligrapher can make errors and have a second or third go at a letter and the unwanted material falls 'on the editor's floor', to recall a dated concept. It now became apparent that our future work would truly be 'multimedia' in that we would include video and sound wherever apposite. The Exeter DVD contains a short film, music and singing, and readings of poems in Old and Modern English. Once you start thinking like this, my job began to expand and I began to think of voiceovers, background music and sets! This will be discussed further in the next section.

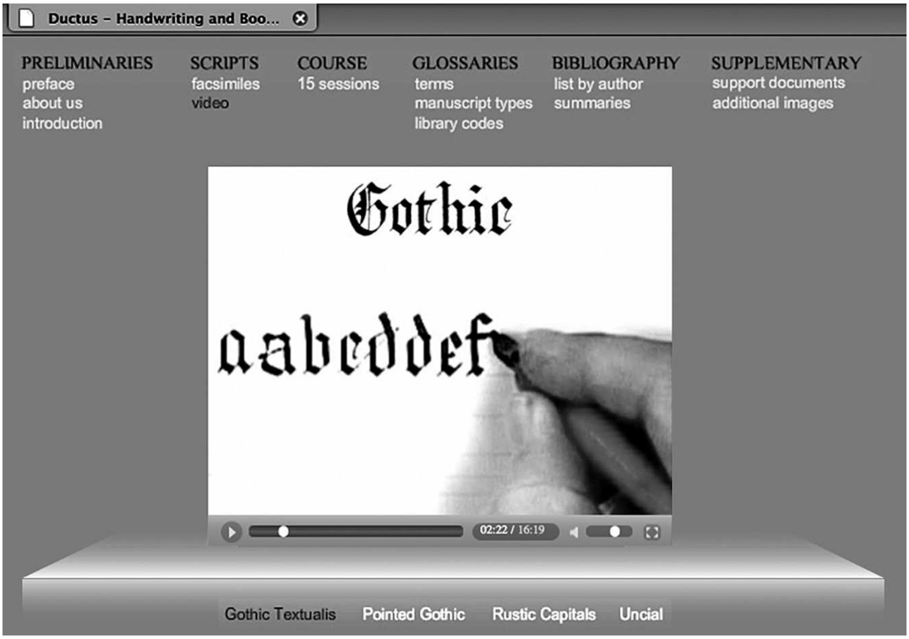


Figure 1. This is a capture of a video showing how a formal Gothic script was written; there are four such videos in Ductus, as can be seen from the labels below the window.

The actual function and use of the program then required consideration. The people I teach using Ductus are usually advanced in their studies and are mature students who are doing the subject in order to acquire an essential skill. This means that an 'honour system' can be used in teaching—the transcription (i.e. answer sheet) is included on the disk so that in theory a student could cheat and get a better mark, but a student who really wants to learn will not do that; they will check their work against the transcription after they have had a go at it. In any event, I do not assess them on how well they transcribed the text. What I am concerned with, and where I think the pedagogical advantage is to be had, is in the students discovering why they made an error and then explaining that to me. This is what I base my assessment on. In any event, if an instructor does not want the students to have access to the transcripts, the software is designed so that we can bar them and make the transcripts only available by a weekly password distributed by the lecturer. Some have chosen this option because their students are younger and not yet ready for the 'honour' system used with advanced students—one has to be certain that they are acquiring their knowledge from their personal endeavours.

Version 2.0 of Ductus is to a certain extent an expansion of the original publication containing more information about a larger number of sample facsimile images, but it has some significant additions. One of these is the inclusion of more video clips about the art of bookmaking and ink preparation. We also added a library of 'Additional Images' arranged by category, containing images of book spines, ownership data relevant for discussion of origin and provenance and coats-of-arms, glossing patterns, damaged manuscripts, papyri, book covers, and diagrams relating to manuscript construction. Flash was used to provide animation and give the program an up-to-date look and feel.

User feedback, gleaned both from student surveys and emails from users around the world, have been unanimously positive and enthusiastic. In some quarters, critics have asked why we charge for our publications and do not just make them accessible to everyone free-of-charge via the internet. The simple answer to this is that what we do is significantly better and that we actually pay libraries for the rights to disseminate their images which, as I have pointed out, are of a much higher quality than what is available at no charge on the web.

Project 2: The Making of a Medieval Manuscript

The second project I would like to say something about is our documentary film, *The Making of a Medieval Manuscript*. This film was made to meet a perceived need in manuscript studies; no one had yet recorded an affordable and detailed scholarly presentation of the complete process of making a book as it would have been done in the Middle Ages and even well into the Age of Print. We chose to make a codex as it would have been done in the thirteenth century in France. I engaged the services of a bookbinder who has been in the business for many years; however, this was the first time that he had ever undertaken to make a manuscript from scratch. Parchment and boards were ordered from Britain and Germany respectively, as were a few other tools required for the job that he did not already have. He already owned the various presses that would be needed for the project.

From the beginning we were concerned that the output be a pedagogically sound tool—students would have to be able to see each step of the process and also be able to examine the internal structure of the codex after it was finished. For this reason, the finished book has cutaway panels on its spine and inside its back covers that allow the students to see the cords attaching the cover boards to the text block, the excavated channels in the boards through which the cords travel, the manner in which the quires or gatherings are stitched to the cords, the way in which the head- and tail-bands are attached, and the ruling technique employed (two different types of ruling are demonstrated—one in lead point and the other scratched into the surface of the parchment with an awl).

The filming was done over three days using two professional cameras. This produced about 40 hours of raw film which in the end was edited down to 40 minutes, just about the right size of a 50-60 minute seminar. We decided that we would make a Powerpoint-style presentation to include on the DVD so that an instructor could take students through each stage of the process at a more generous pace either before or after showing the movie. This also offered us the opportunity to include some 'stills' that dealt with things not seen in detail in the movie. Captions were written for each of the 'slides'; we also broke the process into several stages and included 2-3 screen-fulls of introductory comments at the beginning of each of these.

I did a rough edit of the filmed materials and reduced them to about two hours worth of filming before turning that over to the professional editors. They, of course, knew nothing about what we were doing, so it was important that I give them some guidance in editing the materials lest they inadvertently removed an essential segment (whose absence I may not have noticed until too late). I next wrote a draft of the voiceover for the film before I have seen their first edit. By doing this I felt certain that I had included a description of every aspect of the process, and this would also serve as a further check that they had not left anything essential out. I timed the length of the voiceover to make sure that it was considerably shorter than the movie itself, since I did not want it to sound like I was cramming as much information into 40 minutes as I possible could. Any good teacher knows that a student or viewer can only take in so much information at a time and that they become bored when flooded with facts.

The voiceover was then fitted to each stage of the edited film; there was usually a bit of 'silence' after the voiceover had said what needed to be said at that stage but there was a need for the user to see more of the work being done. And so we introduced some specially-recorded medieval music into these periods of silence, aiming to create a more pleasant learning environment.

The proof is in the pudding. Since the DVD was released last year, hundreds of copies have been purchased by universities all around the world and many commendations have been received. This project leads to the next, which is on the workings of a medieval scriptorium. Here the user will be introduced to every skill required for actually writing out a text and decorating it by hand. There will be demonstrations of how different scripts were written, how scribes used complex abbreviation systems, how they corrected errors, how they made pens from quills, and how inks and paints were made. It will be a multimedia medieval feast. Modest sets have been made devised that recreate the atmosphere of copying out text by both candle- and natural light. And in order to give the users a glimpse into the real world behind these recreations we have filmed some segments in the factory of the pigment maker and the workshop of our modern scribe or calligrapher.

Student reaction has also been very favourable. Most people have no idea how complex a process bookmaking was in the Middle Ages, and few ever stop to consider that each one of the millions of books produced before the Age of Print was made by hand by groups of skilled craftsmen. Few activities would have consumed more resources in the Middle Ages—warfare and cathedral-building spring to mind most readily.

Conclusion

Each of the products we are now producing is designed to allow instructors to use it in a variety of ways reflecting their needs. Careful consideration of pedagogical issues lies at the heart of each DVD and the user interface has to be user-friendly and intuitive, so that complex manuals or sets of instructions are not necessary. Casual 'mousing-over' most labels or buttons should provide the average informed computer user (familiar with internet protocols) with all the information required to explore a program and understand its structure.

Bibliography

Muir, Bernard J. *MS Junius 11.* Bodleian Digital Texts 1. CD ROM. Oxford: Bodleian Library, 2004.

Muir, Bernard J. and Andrew J. Turner. *The Life of St Wilfrid by Edmer of Canterbury.* CD ROM. 2nd ed. Melbourne: Evellum Digital Publications: 2005.

Muir, Bernard J. *The Exeter Anthology of Old English Poetry.* DVD. Exeter: UEP, 2006.

Muir, Bernard J. *The Making of a Medieval Manuscript.* Documentary film. Melbourne: Evellum Digital Publications, 2008.

Muir, Bernard J. Ductus: *Handwriting and Bookmaking in the Middle Ages.* DVD. 2nd ed. Melbourne: Evellum Digital Publications, 2008.

Muir, Bernard J. and Andrew J. Turner eds. *Terence's Six Latin Comedies.* Bodleian Digital Texts 2. DVD. Oxford: Bodleian Library, 2009.

Linking Text and Image with SVG

Hugh A. Cayless

Abstract

Annotation and linking (or referring) have been described as "scholarly primitives", basic methods used in scholarly research and publication of all kinds. The online publication of manuscript images is one basic use case where the need for linking and annotation is very clear. High resolution images are of great use to scholars and transcriptions of texts provide for search and browsing, so the ideal method for the digital publication of manuscript works is the presentation of page images plus a transcription of the text therein. This has become a standard method, but leaves open the questions of how deeply the linkages can be done and how best to handle the annotation of sections of the image. This paper presents a new method (named img2xml) for connecting text and image using an XML-based tracing of the text on the page image. The tracing method was developed as part of a series of experiments in text and image linking beginning in the summer of 2008 and will continue under a grant funded by the National Endowment for the Humanities. It employs Scalable Vector Graphics (SVG) to represent the text in an image of a manuscript page in a referenceable form and enables linking and annotation of the page image in a variety of ways. The paper goes on to discuss the scholarly requirements for tools that will be developed around the tracing method, and explores some of the issues raised by the img2xml method.

Zusammenfassung

Annotation und Referenz sind als geisteswissenschaftliche Stammfunktionen beschrieben worden, die in Forschung und Veröffentlichungen aller Arten verwendet werden. Die Online-Veröffentlichung von Handschriftenbildern ist ein grundlegender Anwendungsfall, in dem der Bedarf für Referenz und Annotation offensichtlich ist. Hochauflösende Bilder sind von großem Nutzen für die Forscher und Transkriptionen ihre Grundlage für Suchen und Stöbern. Damit erweist sich eine Darstellungsweise, in der das Bild der Seite mit seinem Text verknüpft ist, als ideale Form der Publikation von Handschriften. Diese Verknüpfung ist eine Standardmethode geworden; sie lässt jedoch die Frage offen, bis zu welcher Granularität sie erfolgen soll und wie Annotationen von Bildsegmenten erreicht werden können. Der Beitrag stellt nun eine neue Methode (img2xml) vor, mit der Text und Bild XML-basiert verknüpft werden. Diese Methode der »Spurenverfolgung« ist seit Sommer 2008 als Teil einer Reihe von Experimenten

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 145–158.

zur Text-Bild-Verknüpfung entwickelt worden und soll mit Unterstützung des National Endowment for the Humanities fortgesetzt werden. Die Methode verwendet Scalable Vector Graphics (SVG), um den Text auf dem Bild einer Handschriftenseite in einer referenzierbaren Form darzustellen. Sie ermöglicht die Verknüpfung und Annotation des Seitenbildes auf verschiedene Weise. Der Beitrag diskutiert schließlich die wissenschaftlichen Anforderungen an Werkzeuge, die auf Basis dieser Methode entwickelt werden können und untersucht einige der durch die img2xml-Methode aufgeworfenen Fragen.

1 Background

The impetus for the research outlined here comes from my experience working with the online presentation of manuscript texts, and a sense of frustration with the limitations of tools for linking the two. An example of a basic approach to the presentation of manuscript texts may be found in the "First Century of the First State University" (UNC) collection at Documenting the American South (DocSouth). Here, the primary face of a document is its HTML transcription, with links to the XML source, and to medium-quality images of the pages. For any given document, it is possible to see what the original, handwritten texts look like, but the two are not easily viewable side-by-side, and annotations are attached only to the marked-up versions of the text, not the images.

This situation flows naturally from the workflows involved in creating the transcriptions. A scholar will be responsible for transcribing the text, which will then be sent out for encoding in a TEI XML format. After that, a graduate assistant (GA), working with the scholar, will perform quality control and the insertion of notes and links to external materials (including the page images) into the XML text. Finally, the XML is transformed into the HTML version. It is certainly possible to encode links that may be actuated from the images, using image maps for example, but doing so requires the insertion of another step into the workflow and one that is not presently automatable. Given the necessary tradeoffs in time and money involved in this kind of work, such a step is not feasible for most DocSouth projects.

Because we were planning to embark on a new project that would involve the online presentation of a 19th century text, and because I had also been working with transcriptions of papyri destined for presentation alongside images, I began an investigation of what tools were available, and what might be done to automate, and therefore reduce the costs of linking text and image together.

The state of the art for linking manuscript images and texts (both transcriptions and annotations) is to treat the image as a coordinate system and for XML markup in the text to describe a rectangle on the image, containing (for example) a line of text. There are a number of existing tools that provide for user-controlled image annotation. This is typically accomplished by providing the user with drawing tools with which they may draw shape overlays on the image. These overlays can in turn be linked to text annotations entered by the user. This is the way image annotation works on Flickr, for example, and also the Image Markup Tool (IMT). The TEI P5 facsimile markup (TEI P5, chapter 11) conceives of text-image linking in this fashion also. Since these methods require a person to manually create the areas and the linkages between text and image, I began to experiment with methods for producing a representation of the text on a page image that would allow any discrete chunk of text to be referenced (Cayless 2008). This is accomplished via the use of a tool called *potrace* (Selinger) which operates by first converting the source image to a black and white version and then converts the raster image to a vector representation. This vector representation can be output in a variety of formats, including Scalable Vector Graphics (SVG), which is an XML-based format.

The resulting SVG file contains <path> elements, which describe structures in the source image, one path per structure, as cubic Bézier curves. By "structure", I mean a contiguous segment of writing on the page, which might encompass a single letter, a series of connected letters, or other artifact. The fact that each such structure is represented by an XML element means that each one can be referenced and manipulated using standard tools. If the transcription and/or notes are in an XML format also, the image may point at the text. The SVG image, since it is vector-based, is arbitrarily zoomable, but can be based on a coordinate system that will always map back to the source image. The SVG can be overlaid on the source image, parts can be made visible or invisible, transparent, of different colors, rotated, etc. An additional benefit of producing a vector representation of the page image is that it becomes quite easy to detect some aspects of page structure, like lines.

The proof-of-concept work on this method was successful, and a fully Open Source toolchain (provisionally called *img2xml*) was developed that started with a page image from a letter, with an existing transcription marked up in TEI XML, and resulting in a web view where line of text were linked to lines in the image, and vice-versa. Subsequently, the Carolina Digital Library and Archives submitted a National Endowment for the Humanities (NEH) Startup Grant to begin developing the experimental version into a functional prototype. The proposal has been funded, and development of the prototype will begin in mid-2009.

1.1 Methods

The methods and a summary of the code that has been developed thus far are outlined below. Since the project is at an early stage, and further development is anticipated under the recently-awarded NEH Startup Grant, the codebase is expected to evolve quickly. Updates will be posted on the project's website..

The *img2xml* process may be broken down into the following steps: pre-processing of images, generation of the SVG tracing, post-processing of the SVG, analysis of the

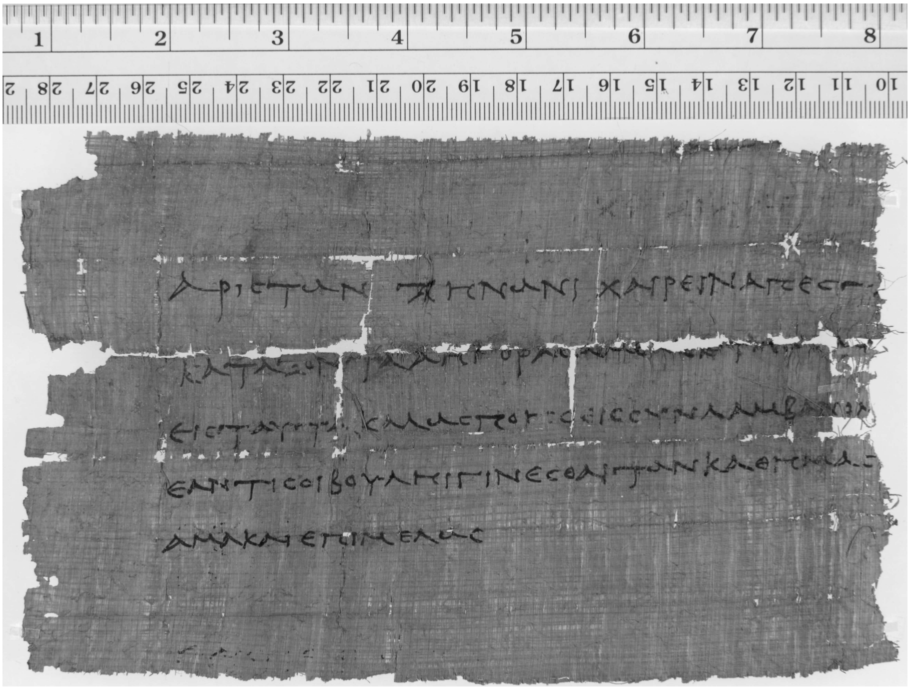


Figure 1. Fragment of a letter from Ariston to Zenon (P. Mich I 78).

SVG, and presentation. Pre-processing would involve filtering the image to whatever extent possible to improve the chances of *potrace* generating an SVG as free as possible of artifacts that are not letters. Since *potrace* relies on first flattening the image's color space to a single bit (black or white), it depends on a black or white threshold being set. During processing, a decision is made, based on the darkness of a given pixel, whether it should become black or white. With materials like papyrus, where the support itself is quite dark, setting this threshold on an unprocessed image (see Figure 1) can involve a good deal of trial and error. But since the support and the ink are different colors, a tool like *ImageMagick* can be used to delete the support from the image, leaving the ink behind (see Figure 2). This technique is also useful for images where there are blemishes (as long as they differ in color from the ink itself).

Once the tracing program and any post-tracing cleanup (including the conversion of relative paths to absolute, and the addition of id attributes to the path elements) have run, the SVG file is ready for analysis. The initial proof-of-concept system exper-

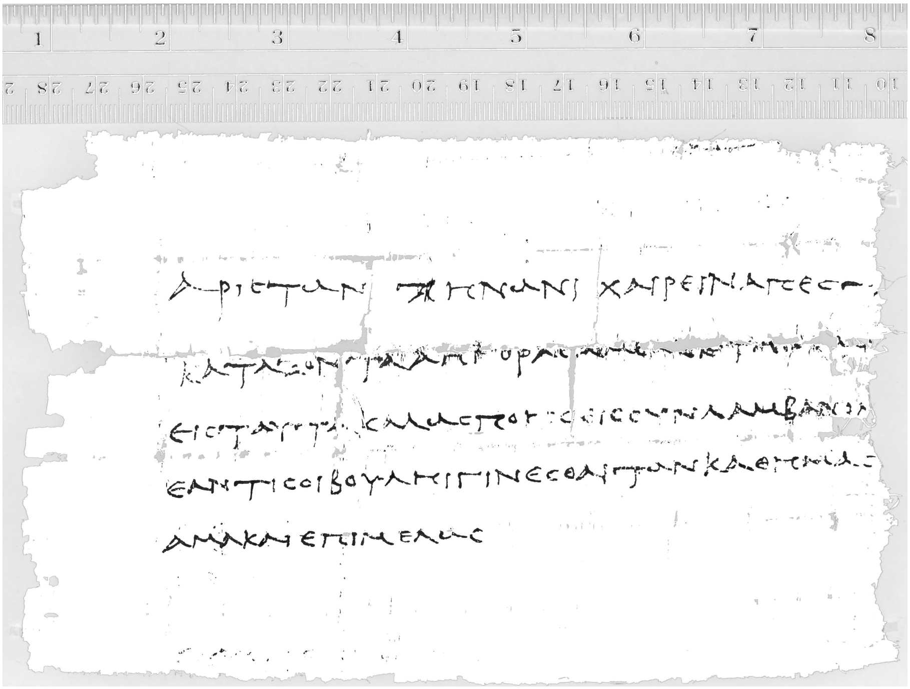
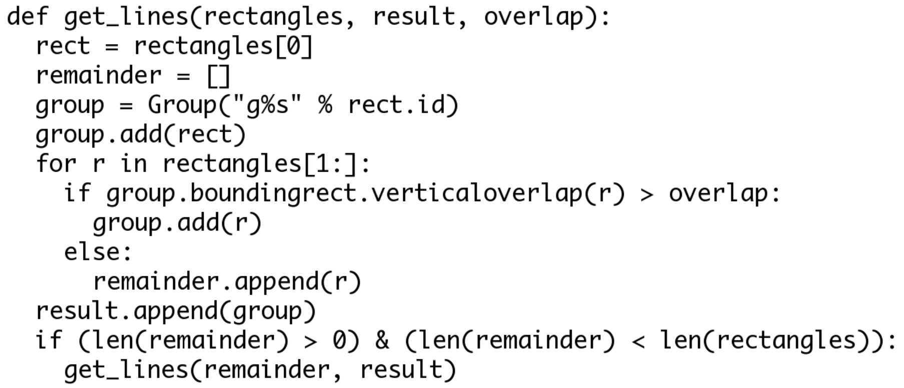


Figure 2. Fig. 1 processed to remove the color of the papyrus.

imented with very simple structure detection. Lines are detected by parsing the SVG file and loading the paths into Python data structures containing the points denoted in the SVG paths. Then, for each polygon, a bounding rectangle is determined by finding the outermost top, bottom, left, and right coordinates of the polygon. Since the original shapes are not polygons, but Bézier curves, there exists the possibility that the control points (which determine the shape of a curve, but do not lie upon that curve) will result in a rectangle that is too large or too small, but thus far in practice this does not appear to be a significant issue. As the rectangles are being extracted, each one is appended to an array, and the area of each rectangle is also appended to an array. The rectangles areas constitute a rough relative measure of size, and this measure is then used to prune objects that are unusually small or large. This helps remove artifacts such as dust spots that should be ignored as components of the document's structure. It can also be used to dispose of the outlines of pages that have been scanned against a black background.



```
def get_lines(rectangles, result, overlap):
  rect = rectangles[0]
  remainder = []
  group = Group("g%s" % rect.id)
  group.add(rect)
  for r in rectangles[1:]:
    if group.boundingrect.verticaloverlap(r) > overlap:
      group.add(r)
    else:
      remainder.append(r)
  result.append(group)
  if (len(remainder) > 0) & (len(remainder) < len(rectangles)):
    get_lines(remainder, result)
```

Figure 3. Line extraction code in Python.

Removing structures with areas smaller than two standard deviations below the mean seems to produce useful results.

After filtering, the rectangles are sorted on first the x then the y axis, which improves the efficiency of the analyses to follow. The program proceeds with line detection. During this process, shapes are added to group objects, each of which contains an array of shapes and a bounding rectangle. At the end of the analysis, these will be serialized into the output SVG as <g> elements.

The algorithm for sorting into lines is shown in Figure 3. The method starts with the rectangle at the beginning of the sorted array (which will be the topmost, leftmost shape), adds it to a group object, and then cycles through the rectangles array, looking for rectangles that overlap more than a defined percentage (50% is a good number) with the group. As each new rectangle is added, the size of the group's bounding rectangle grows. If the next rectangle in the array does not overlap, it is added to a remainder array. As long as the remainder still has shapes in it, the method is called recursively on the remainder. At the end of the process, all rectangles will have been organized into line groups.

Word detection is a process with which the project has only begun to experiment. In texts which exhibit significant space between words this should be possible using processes similar to those used in Optical Character Recognition (OCR) programs. In texts like the papyrus example from Figure 1, however, this will not work because the text is written in *scriptio continua.* If a transcription already exists, then that could be used to help align the shapes traced from the image with itself. This raises the question of how *img2xml* fits into the workflow of digitization to online production, and how it might relate to manuscript OCR.

Manuscript OCR is a problem without a satisfactory solution to date, though the technology is improving all the time (see, for example Gilbert and Roland Tomasi's article in this volume). The main difficulty in manuscript OCR is the lack of consistency in handwritten versus printed text. Progress can be made on relatively uniform hands with a significant amount of work in training the recognition algorithms, but the cost/benefit ratio is a problem, and this is compounded when one is faced with a situation like Greek papyri, where there are thousands of short texts, in thousands of hands. The process used by *img2xml* does share some workflow steps with OCR, which does the same sort of image downsampling prior to isolating and attempting to identify symbols and structure on the page. But *img2xml* is envisioned as complementary to OCR, and will be useful even when (or if) fully-functional manuscript OCR is a reality. *Img2xml* at its core is a tool that produces a vector representation of text, which can then be manipulated and linked. It does not necessarily attempt to aid in the production of a transcription of the text it represents, though a scholar producing such a transcription could use the SVG tracing as a reference. The tool's main payoff is likely to be on the publication and presentation side.

1.2 Further Work

The award of the proposed NEH Startup grant was announced on March 9th, 2009. This award will enable the project to proceed at an accelerated rate. It will focus on developing a system to present the diary of a 19th-century UNC Chapel Hill student, James Dusenbery. This manuscript has already been digitized and the transcription encoded according to the TEI guidelines.

A graduate assistant will be funded as part of the grant budget who will be responsible both for enhancing the previously-encoded Dusenbery Diary by adding tagging to denote the lines in the transcription and by converting the text to TEI P5, and for testing and evaluating the tools developed for viewing and working with the images and text. In addition, the grant will help fund some programmer time to further develop the tools we have begun working on. We plan to complete the following steps as part of the project[1]:

1. Potrace needs to be tested on a wider variety of manuscript images in order to determine the best practices for using it with these types of image.
2. We will develop methods for preprocessing images and detecting the ideal white/black cutoff to produce an optimal SVG tracing. Or, failing that, develop a web interface whereby users can run traces via a web interface and choose the best one.
3. The automated line detection routine developed for the proof-of-concept tool

[1] This list is adapted from the proposed workplan for the NEH grant.

works well for relatively horizontal, left-to-right writing, but the algorithm will need further development in order to handle text that runs (for example) at an angle. Various examples of nonstandard texts will be selected from our collection and the tool will be tested against these to determine its effectiveness and what additional development may be needed.

4. The automated detection of word boundaries will be evaluated as well, again by testing the analysis tool against a variety of images. Further development will be undertaken on the use of the transcription as a means for detecting word boundaries where they do not exist on the page.
5. OpenLayers was patched during the proof-of-concept work to allow the representation of complex paths and their serialization as SVG. This patch requires substantial improvement in order to make better use of OpenLayers' zooming capabilities. The work done to date will be refined and tested. Upon completion, the patch will be submitted to the OpenLayers developers for possible inclusion in the library.
6. Work will be done on the web interface to allow for simultaneous paging through page images and transcriptions/annotations.
7. A graduate assistant will work on the updates to the previously digitized Dusenbery text detailed above.
8. The project will be published as a collection within the Documenting the American South program.
9. All of the code developed for the project will be published in an open code repository along with documentation of the tools and a user manual. We will present the results (as well as ongoing work) at one or more of the major Digital Humanities and/or Digital Library conferences.

The interface that is developed will demonstrate features like links from notes to text in the image, the ability to link or pan/zoom to any part of the text in the image from the transcription, highlighting text search results on the surface of the image, and the marking of editorial emendations, such as expanded abbreviations and other types of editorial addition or deletion on the image itself.

2 Scholarly Requirements for Operating on Text and Image

In designing the presentation system for the Dusenbery Diary, I have made several assumptions about the functionality that should be present in the outcome, which I hope will serve as a prototype for other online manuscript publications. Some of these assumptions about desirable requirements for a system to link text and image which underlay the efforts to date are that:

1. the software should enable bidirectional linking (that is, both text to image and image to text).

2. it should be possible to link at the level of any structure on the page image.
3. images should be as manipulable as possible (i.e. panning, zooming, rotation).
4. the results should be presented in an online (browser-based) form.
5. any systems or methods constructed should be based on Open Source tools and released under an open license.

Assumptions like these should always be examined. That linking both from text to image and from image to text is a desideratum seems obvious. In an interface that presents both, either one should be able to become the primary focus of the reader's attention, but it should always be possible to call up the other and smoothly move between them.

The question of what level to link at is more complicated. A straightforward answer is that linking should be as granular as possible, but it is also true that the line is a basic unit of reference (like the page) and that line-level linking is a basic requirement. Word- and letter-level linking are also important desiderata. A prerequisite for either, however, is the detection of all written symbols on the page, which can be then grouped into words, lines, and so on. Of course, word detection is much easier when letters are organized into words, with spaces in between, which is not the case in most ancient texts. This detection of document structure is the problem that OCR attempts to solve, by first matching symbols to letters and then attempting to recognize words from collections of symbols. As I noted above, the experiments to date have proceeded by doing only structure detection which does not attempt to recognize symbols. An OCR process could be combined with *img2xml* in the future, which would enable superior structure detection. At this time, *img2xml* method enables linking at the level of the symbol, and at the level of any larger structures that can be detected based on the arrangement of the symbols.

Assumptions 3 and 4 are related, since the platform dictates to a large extent the capabilities of the image viewer. Fortunately, modern browsers are highly capable in this regard.[2] Provided that the page images have been digitized at a high enough resolution, there are a number of techniques that provide for panning, zooming, and rotation. The original demonstration version of the *img2xml* viewing tool employed a Javascript mapping library called OpenLayers, which supports pan-and-zoom functionality in a variety of ways. The author has successfully linked it to aDORe djatoka (Chute), an Open Source image server that uses JPEG2000 as the service format. Experimentation to date has focused entirely on Javascript-based methods and avoided reliance on proprietary technologies such as Flash. To date, this approach has not met with any insoluble obstacles.

Finally, it was the author's conviction from the beginning that all of the tooling and all of the methods developed during this project should be released under open

[2] The main difficulty being that Internet Explorer (as of IE 7.0) does not have native SVG support.

licenses. Proprietary technologies certainly have their place, but the kind of work under development is being designed for the online publication of scholarly editions. For this kind of work, there is unlikely to be much monetary reward and the long-term survivability of the output is of primary concern. Editions produced using tools like the ones the *img2xml* project will build may not work in their current form 20 years after their release, but if the formats and code are open, it will at least be possible to migrate them to new, working forms. The use of proprietary formats and closed software makes indefinite survival a great deal less likely.

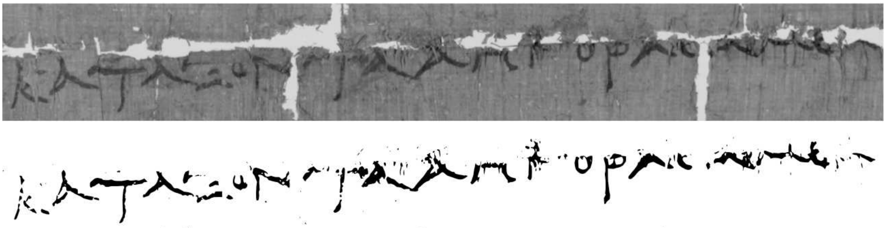
3 Problems and Solutions

In examining ways one might link between an XML transcription of the text and an XML overlay of the text, one quickly runs into problems involving overlapping hierarchies: single paths may include multiple letters or words, for example, and there may be single letters constituted from paths. As I noted above, the process of generating the SVG tracing involves the conversion of the image to a black and white (1-bit) bitmap, wherein each pixel is either 0 or 1. This makes it possible for the software to reproduce the shapes in the original source in vector format, but it also means the originally layered text has been flattened. While it might have been clear that the stroke of one letter runs over the top of a second in a color image, that layering is lost in the SVG, and the two letters are a single shape in the output. Such issues of information loss complicate the detection of structure in documents.

The figure below (derived from the papyrus in Figure 1) highlights some of these issues. Notice that the initial kappa is represented by no less than eight paths, while part of the downward stroke of the final alpha in κατἀξοντα connects to the following word, ἅ.

A variety of possible solutions to this problem presents itself, including editing the SVG so that the single path is split into two, or indicating word boundaries by drawing boxes that may intersect with paths. Even more pronounced examples of the layering problem are not hard to find: the Archimedes Palimpsest (Noel) contains a wealth of them, and highlights the value of image processing as a tool to enable scholars to read difficult texts. A palimpsest by definition contains at least two layers of text, and typically it is the obscured one that holds more scholarly interest (see Figure 5). *Img2xml* itself does not hold out any promise as a tool for separating layers—for that one needs multispectral imaging of the kind used on Archimedes—but once those layers can be teased apart, it could serve as a means to highlight the undertext on any image, or to demonstrate the interaction between layers.

The potential capabilities of this method quickly reveal some of the shortcomings in our representational formats when one begins to explore ways of linking between



κατάξοντα ἃ ἠγοράσαμεν

ΚΑΤΑΞΟΝΤΑΑΗΓΟΡΑΣΑΜΕΝ

Figure 4. Text and transcriptions of P. Mich. I 78, line 2.

tracings, images, annotations, and transcriptions. Pure linking is simple: all that is required when the formats are XML-based is that elements have unique identifiers. The SVG is a derivative of, and provides a usable coordinate system for, the source image. If one is careful (as for example the Archimedes Palimpsest project has been) to keep multiple images in different lighting aligned, then the SVG can serve as a map for all of them. But the semantics of SVG are almost purely geometric. It encodes shapes, with additional support for links, text, embedded images, and animation. This means there is no inherent way to express the significance of a grouping or a feature in SVG, nor the relationships between them. So, for example, we may consider how one might indicate that a set of paths in the SVG document signifies a word in the transcription. Assuming this is the uncomplicated case, where the shapes on the page do not somehow join members of one word to the next, we can combine those paths into an SVG group (<g>) element or we can draw a box within whose bounds all of the paths fall. The grouping at least has the advantage of establishing a parent-child relationship between the group and its members, and the <g> element, given an id-attribute, can be referenced externally or internally. Unfortunately, this method will only work in the simple case where a structure on the page does not contain letters from two different words. Unless a workaround can be found (and some possibilities will be discussed below), the creation of a container outside the hierarchy of paths and groups will be necessary. It is easy to create a rectangle which when rendered will contain an arbitrary set of paths, but because of SVG's geometric semantics, there will be no obvious way to

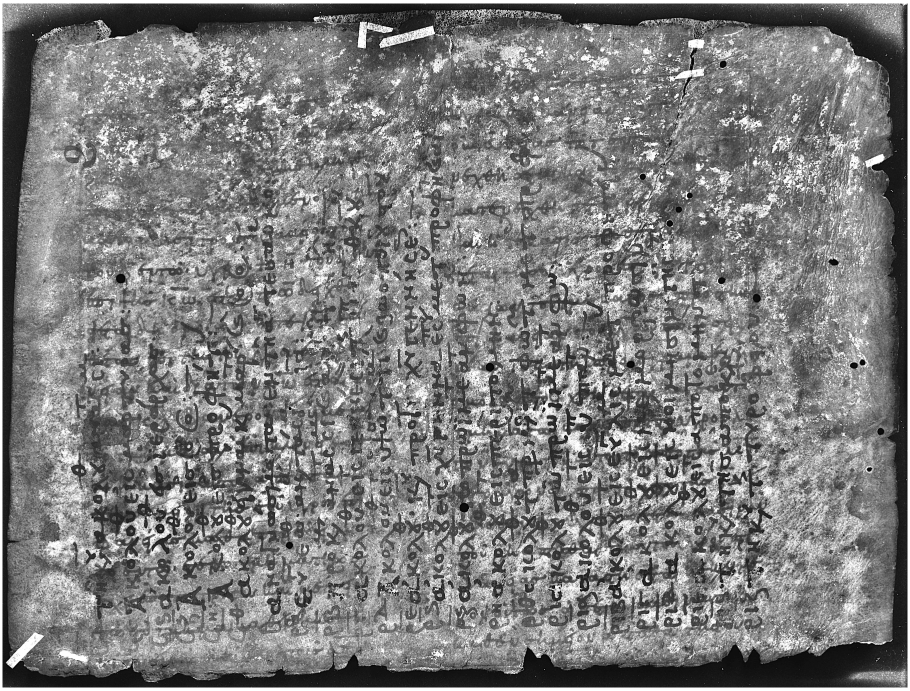


Figure 5. Archimedes *On Floating Bodies* folio 13v—under ultraviolet light.

indicate the relationship between that rectangle and the paths it "contains."[3] We are left then with a need either to work around the problem of overlapping hierarchies so that we can exploit the parent-child relationship established by <g> or to produce a method for defining relationships on the order of "isMemberOf" so that elements connected only by geometry may be semantically linked.

There are a variety of possible solutions to the issues presented by the hierarchical approach. For example, the paths forming κατάξοντα ἅ from Figure 3 could be split apart using a process that could find the intersection points of a word-dividing line (e.g. from a rectangle containing the word κατάξοντα) and the path representing the two letters alpha, and then split that path into two derivative paths, each of which would be associated with a different word and could become children of a <g> connected to a word in the transcription. This method would avoid damage to the actual tracing while

[3] I say "contains" in quotes, because the containment is only apparent when the document is rendered visually.

allowing the types of reference that are likely to be useful. The derivative paths could be placed in the same document and only activated as needed. But such a method would immediately raise the need for semantics again. We would want to be able to discover the relationship between the original, connected "…α ἅ" and its divided members since there is no inherent way to distinguish between one <path> element and another.

Regardless of the method used to associate paths and partial paths with words, letters, or notes, then, it will be a requirement that semantics be added to the SVG document somehow. Some possibilities for adding semantics to SVG documents include embedding relationship metadata using SVG's <metadata> element or developing a microformat (Microformats), perhaps depending on the class attribute, which is available on all displayable elements (Schepers). A further possibility would be to create an external document using the Resource Description Framework (RDF) that defined the relationships between elements within the SVG. This would have the further advantage that it could be used as the means to link the transcription and notes to the SVG as well, using a language with richer semantics than a plain URI.

4 Conclusions

The range of possibilities presented by the experimentation done to date on this method demonstrate its potential as a tool for the presentation of scholarly research on digitized manuscripts. Details of the implementation, particularly as regards the methods used to link between SVG tracing, transcriptions, and notes, remain to be worked out. There are also a range of issues, such as the layering problem and the need for semantic linking that must be explored. But there are a range of potential solutions to these problems which will be explored as part of the NEH Grant. The progress of the project may be observed at the project repository, where code and documentation will be released as they are produced.

Bibliography

Cayless, Hugh. "Linking Page Images to Transcriptions with SVG." *Proceedings of Balisage: The Markup Conference (2008).* <http://www.balisage.net/Proceedings/html/2008/Cayless01/Balisage2008-Cayless01.html>.

Cayless, Hugh. "*Img2XML* project repository." <http://github.com/hcayless/img2xml/tree/master>.

Chute, Ryan. *aDORe djatoka.* <http://african.lanl.gov/aDORe/projects/djatoka/index.html>.

Documenting the American South (DocSouth). <http://docsouth.unc.edu>.

The First Century of the First State University (UNC). <http://docsouth.unc.edu/unc>.

Flickr. <http://flickr.com>.

Holmes, Martin. *The Image Markup Tool.* <http://tapor.uvic.ca/~mholmes/image_markup/>.

ImageMagick. <http://www.imagemagick.org/index.php>.

Microformats. <http://microformats.org>.

Noel, William, et al. *The Archimedes Palimpsest.* <http://archimedespalimpsest.org/index.html>.

OpenLayers, v. 2.7. <http://openlayers.org>. (Release notes for v. 2.7 at http://trac.openlayers.org/wiki/Release/2.7/Notes).

Schepers, Doug. *Reinventing Fire* 'Blog Archive' SVG Text, Semantics, and Accessibility. November 7th, 2006 at 5:24 am. <http://schepers.cc/?p=11>.

Selinger, Peter. *Potrace, v. 1.8.* <http://potrace.sourceforge.net>.

Selinger, Peter. *Potrace: a polygon-based tracing algorithm.* <http://potrace.sourceforge.net/potrace.pdf>.

TEI Consortium. *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Ed. Lou Burnard and Syd Bauman, 2008. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html>.

World Wide Web Consortium. *Resource Description Framework (RDF).* <http://www.w3.org/RDF/>.

World Wide Web Consortium. *Scalable Vector Graphics (SVG).* <http://www.w3.org/Graphics/SVG/>.

The Ghost in the Manuscript: Hyperspectral Text Recovery and Segmentation

Patrick Shiel, Malte Rehbein, John Keating

Abstract

Major activities in palaeographic and manuscript studies include the recovery of illegible or deleted text, the minute analyses of scribal hands, the identification of inks, and the segmentation and dating of text. This article describes how Hyperspectral Imaging (HSI) can be used to perform quality text recovery, segmentation and dating of historical documents. It provides a comprehensive overview of HSI, and associated computational and segmentation techniques used for two experimental investigations: (i) a 16th century pastedown cover, and (ii) a multi-ink example typical of that found in, for example, late medieval administrative texts such as Göttingen's *kundige bok*.

Zusammenfassung

Paläographische Forschung und das Studium von Handschriften umfassen das Wiederherstellen unlesbaren oder getilgten Texts, die genaue Analyse von Händen, die Identifizierung von Tinten sowie die Segmentierung und Datierung von Text. Dieser Beitrag beschreibt, wie hyperspektrale Bildverarbeitung (*Hyperspectral Imaging, HSI*) für diese Aufgaben angewendet werden kann. Er liefert einen umfassenden Überblick über die Technologie und beschreibt die rechnergestützten Verfahren und Methoden der Segmentierung, die experimentell für zwei Fallstudien entwickelt wurden: (i) zur Untersuchung eines Bucheinbandes aus dem 16. Jahrhundert und (ii) zur Analyse der Gebrauchsspuren verschiedener Tinten, wie sie etwa typisch für spätmittelalterliche Amtsbücher sind.

1 Introduction

The condition of medieval manuscripts ranges from those that are fully legible to those that can only be read in part, and their legibility is determined by the manner in which they were preserved and treated throughout the ages. In some cases deterioration is due to processes such as fading or staining; in others, the text may have been interfered with in some way. For instance, in the Irish context, the oldest (12th century) surviving

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 159–174.

manuscript written entirely in Irish, *Leabhar na hUidhre* (*Book of the Dun Cow*) was subject to part-erasure and rewriting by a scribe who was active at some point between the 12th and the 14th centuries (Ó Cróinín). In the German context, we refer to a 15th century manuscript (*kundige bok*) containing a legal text that is characterised by many revisions over a period of approximately 50 years (Rehbein). It is a multi-layered text of 330 paper pages, with its different layers representing the various stages in the development of the town law over the years. Revealing its layers is crucial for understanding the text and for historical studies.

For these examples, it is not only necessary to identify the different scribes in the manuscript but also to deal with the issue of the same scribe writing at different points in time. The work on the digital edition of *kundige bok*, for example, which aims at visualising the textual evolution, has so far relied mainly on two techniques: (i) looking for dated entries and contextual (internal and external) information, and (ii) using palaeographic, codicological and linguistic methods to analyse a range of identifying factors. These would include analysis of the scribes' style of writing over the years (i.e. the writing process itself and customs in using certain words and phrases), inks (as far as the human eye can distinguish the colours), and the paper the text was written on (analysis of watermarks etc.), etc. (Bischoff).

Using this approach, it was possible to assign slightly more than half of the textual alterations to text layers (i.e. stages of the town law) and to bring them into chronological order. This required, however, a lot of experience; it took time to familiarise with scribes and scribal habits of that particular place and time, and objectivity in one's decision could not always be assured. Overall, it was a time consuming process, leaving behind a good amount of uncertainty. There are still more than 40% of the medieval scribes' changes to the text that cannot be assigned to a text layer or brought into chronological order in a satisfying way at all. The digital edition of *kundige bok* copes with this by ensuring transparency in the editorial method, making decisions and uncertainty visible to the users and allowing them to dynamically create text layers on their own using the facsimile provided alongside the transcriptions. But shall this be the end of our efforts? Uncertainty in revealing the textual evolution grows when: (i) too many entries in a particular context are not dated, (ii) the same scribe makes corrections at the same passages but (likely) at different times, (iii) ink colours are too similar for the human eye to distinguish, (iv) entries are too short to survey scriptural characteristics or do not even consist of text at all (e.g strike-throughs).

Two central research questions related to the uncertainty of textual development, of interest to all scholars, therefore, are:

- To what extent is text recovery (e.g. in the cases of palimpsest, fading, deliberate removal, etc.) in key medieval texts possible using current technologies?

- To what extent is it possible to establish irrefutable scientific evidence for interpretation of questioned documents, e.g., identify the different hands (inks)?

In this article, we illustrate how to provide answers to these questions using modern scientific techniques and emerging forensic technology, i.e. hyperspectral imaging (Chang 2003) and associated image processing techniques (Chang 2007). In particular, we are interested in the application of hyperspectral segmentation techniques to multi-layered manuscripts to help solve these problems and overcome the uncertainty of the textual development. It focuses thereby on working particularly on the dating issues of the later added entries, thus allowing us to bring them at least into a chronological order where palaeographic means alone would fail.

The availability of a hyperspectral scanner, a Forensic XP-4010, has presented the authors with opportunities to subject damaged or illegible texts to a modern scientific re-examination. The scanner has the potential to read various different layers of a manuscript in a manner not possible to the human eye and to analyse elements of its composition. As such it presents the possibility of retrieving text that has been lost through fading, staining, overwriting or other forms of erasure. In addition, it offers the prospect of distinguishing different ink-types, and furnishing us with details of the manuscript's composition, all of which are refinements that can be used to answer questions about date and provenance. This process marks a new departure for the study of manuscripts, for the authors, and may provide answers to many long-standing questions posed by palaeographers and by scholars in a variety of disciplines. Furthermore, through text retrieval, it holds out the prospect of adding considerably to the existing corpus of texts and to providing many new research opportunities for coming generations of scholars. In this introductory chapter on hyperspectral imaging, we concentrate on two key processes: text recovery and text segmentation.

2 Background and Methodology

The investigative and analytical methods described here are based on a novel and highly specialised technique called Hyperspectral Imaging (HSI), and sometimes referred to as Optical Reflectance Imaging. HSI is a non-destructive optical technique that measures reflectance (fraction of light reflected) characteristics of a document with high spatial and spectral resolution. An HSI device, operating as a reflectance spectrometer, records a sequence (typically hundreds) of digital images of the selected manuscript area (with maximum dimensions 50mm x 50mm) illuminated with monochromatic light from a tunable light source from 350nm (near-UV) through the entire visible range and up to 2400nm (infrared). The value of each image pixel in the recorded image sequence represents an accurate measurement of the reflectance curve for a tiny—13 micron square—area on the document. Analysis of all spectral curves, essentially a cube of information, provides information about the physical characteristics of questioned manuscripts.

HSI, together with modern two-dimensional spectrum software and three-dimensional image and visualisation software, provides modern researchers working in the field of historic documents analysis with opportunities for forensic examination that were heretofore unavailable. Methodologically, there are two main fields of applications of this technique: (i) the extraction of relevant historic, diplomatic and palaeographic information from documents, and (ii) the investigation of the impact of environmental conditions on document condition and of degradation effects on writing materials and substrates. In particular, reflectance curves found in different sections of the manuscripts can be compared with each other in order to determine whether different types of inks had been used during text composition or to identify modifications that occurred during the manuscripts' history. Light spectroscopy analyses may also be conducted to aid recovery and segmentation. Fluorescence occurs when an object emits a high wavelength (low energy light) following illumination by a shorter wavelength (higher energy light) due to molecular absorption of part of the incident light. Furthermore, the spectral curves may be compared with those in international databases containing typical ink spectra to determine and date the kind of ink or pigment used. The image cube recorded using the technique may be used to enhance the visibility of hidden material such as palimpsest or erased text.

This methodology for manuscript analysis is of significant interest to archivists, conservationists, and scientists interested in non-destructive historical document analysis. Klein et al. (2008) recently provided an excellent description of the basic concepts, working principles, construction and performance of a HSI device specifically developed for the analysis of historical documents. Their custom-developed quantitative hyperspectral imager is currently used by the *Nationaal Archief* (National Archives of The Netherlands) to study degradation effects of artificial samples and original documents, exposed in their permanent exhibition area or stored in their deposit rooms. Earlier Klein et al. used their device to record the variation of spectral reflectance on a historic 17th century map, and also used the instrument to compare the local variation of the yellowness index of reference papers stored in a bound volume, and loose sheets. Using HSI they determined that yellowness occurs within 20mm border in both cases, and that the yellowness index was much higher for the bound paper than the loose paper. Ongoing HSI investigations allow researchers to detect and visualize differences in aging processes on a document, and are particularly useful when taken, for example, before and after an exhibition, whereby it is possible to investigate the effects of exhibiting and handling on document yellowing (Padoan et al.).

Padoan et al. have conducted HSI supported palaeographic investigations of the first manuscript written in Dutch (14th century), held by *Koninklijke Bibliothek* (Royal Library of the Netherlands). In particular, the investigation centred on whether a coat of arms, drawn on the lower part of the first page, could have been ascribed to the famous 15th century Flemish bibliophile *Lodewijk van Gruuthuse* or whether the element may have been added at the same time as the 17th century text surrounding it (Padoan et al.). They found that the HSI derived image shows no correlation between the coat-of-arms and the surrounding text, while strong similarities were observed within areas where corrections were made in the 15th century.

There have also been several other recent reports of promising results from the application of the HSI analysis of paintings and conservation of works of art (Fischer and Kakoulli), paper discoloration and foxing (Missori et al.), and detection of iron-gall inks (Havermans). It is our observation that the successful application of HSI in palaeographic and codicological studies emerges from equitable partnerships of researchers and scholars in humanities, computing, and natural sciences.

3 Reflectance and Fluorescence Spectroscopy

Light spectroscopy is the study of light that is emitted by or reflected from objects. When applied to hyperspectral image analysis, spectroscopy deals with capture and examination of images using a large portion of the light spectrum. Light and other forms of electromagnetic radiation are commonly expressed in terms of their wavelength; each photon of light has a wavelength determined by its energy level. When a single frequency light wave comes in contact with an object several different phenomena may be observed depending on the object's composition and wavelength of the incident radiation, for example, the incident light may be reflected, absorbed and turned to heat, or transmitted (and refracted). Single frequency incident light is uncommon, however. Normally, visible light striking an object contains many frequencies, and when this occurs the object will selectively reflect, absorb or transmit certain frequencies. Some excellent introductions to the various aspects of light spectroscopy are given by Hapke who discusses, in detail, the theory of reflectance and emittance spectroscopy, and Ashutosh and Schulman who comprehensively describe fluorescence spectroscopy in their volume.

HSI analysis, therefore, refers to the analysis of spectral images taken at a sequence of spatially aligned wavelength bands. Simple spectral imaging systems, such as a digital camera acquire intensity images in 3 bands (red, green and blue). Hyperspectral Image systems typically image a scene in hundreds of indexed bands, utilising information from a wide region of the electromagnetic spectrum. Modern hyperspectral image sensors can be used to capture the attributes of light emitted by materials, and its variation in energy with wavelength, at a series of narrow and contiguous wavelength bands. For our investigations, we used the Forensic XP-4010 forensic document examination system (from MS-Macrosystems, The Netherlands) to acquire high-resolution optical

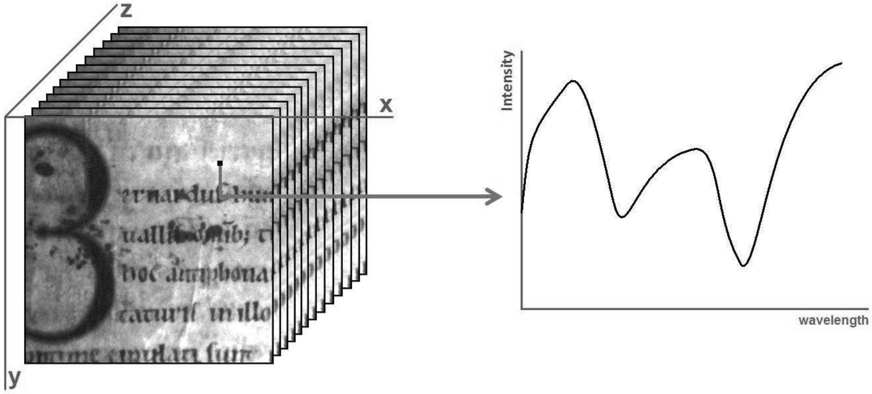


Figure 1. A Hyperspectral cube (left); hundreds of spatially recorded images acquired contiguously over a wavelength region. The reflected energy spectrum for an individual pixel is shown on the right.

and infrared images. The sensor measures reflectance, absorption, transmittance and fluorescence spectral information in the ultraviolet, and 400–1000nm region. The data obtained is in the form of a data-cube (Figure 1) which represents the image information as a data set in three dimensions; two of the data cube's axes represent the spatial data, while its third axis represents the spectral information. This image cube consists of hundreds of spatially recorded images, acquired contiguously over the wavelength region. Each pixel within a hyperspectral image-cube represents the reflected energy spectrum of materials spatially covered by the pixel.

Spectral reflectance, the fractional amount of incident energy that is reflected from a surface with respect to wavelength, is one of the fundamental attributes obtained when performing hyperspectral analysis of documents. In general, an object's reflectance varies with wavelength, as incident energy at particular wavelengths is absorbed or scattered in different directions. Materials that have a similar reflectance under natural light may have vastly different reflectance under light of a specific wavelength from other regions of the light spectrum. The importance of this, in the context of analysis of documents, is that this difference in spectral reflectance allows for classification of different features of the document which, under natural light, look identical. Furthermore, the spectral reflectance also depends on the orientation of the object surface which can be a problem with items that are deformable, such as manuscripts. This may be overcome, however, by implementing diffuse illumination within a light-proof observation chamber. In an experimental investigation on text recovery, detailed below, we demonstrate how spectral reflectance may be used to segment text written with three different pens that appear similar to the human eye.

Luminescence is a phenomenon where an object emits light due to chemical reaction, electrical energy, subatomic motions, or stress. A particular type of luminescence, called fluorescence, is of particular interest when performing hyperspectral analysis of documents. Fluorescence occurs when an object emits a high wavelength (low energy light) following illumination by a shorter wavelength (higher energy light) due to molecular absorption of part of the incident light. The object absorbs part of the incident light which causes excitement of the molecules within the object and then emits energy in the form of lower energy light usually within the visible range of the spectrum. In an example, later in this chapter, we show how the application fluorescence spectroscopy can be useful in recovering unreadable text from historical documents in our sample 16th century book cover.

Tilley provides an excellent discussion on the relationship between light, the optical properties of materials and colour, and is particularly useful for those interested in using HSI to distinguish different materials, for example, inks on paper. HSI observations of historical documents written on paper, parchment, etc. are complex spectral combinations of the reflectance of a collection of materials that have different temporal degradation properties. Vaarasalo provides a useful discussion on the optical properties of paper, which loses its optical properties as time passes (van der Reyden). More recently, de la Rosa and Bautista have discovered that to find the presence and concentration of different colorants or components in the paper, it is only necessary to know the spectra and fluorescence lifetimes (at 337.1nm). They indicate that these kinds of measurements could be useful for studying the paper's long-term stability and how aging affects it, and is particularly important in the preservation of paper-based historical records (Committee on Preservation of Historical Records). Light spectroscopy has also been particularly useful in the examination of paper aging, for example, in accelerated aging experiments (Bansa) and on the evaluation of whiteness and yellowness (Smith).

4 Experimental Investigation: Recovery of Hidden Text

We are especially interested in the use of hyperspectral analysis to support the recovery of hidden text, and in particular, the kind of recovery that requires substantial conservation efforts, or the disassembly of manuscripts and their bindings. We believe that it is possible to employ non-invasive and non-destructive investigative techniques to search for hidden text prior to embarking on physical analyses and treatment of historical texts.

A particularly interesting example of this high-precision, skilled and manual work is reported by Quandt who conducted a detailed physical analysis and treatment of a late

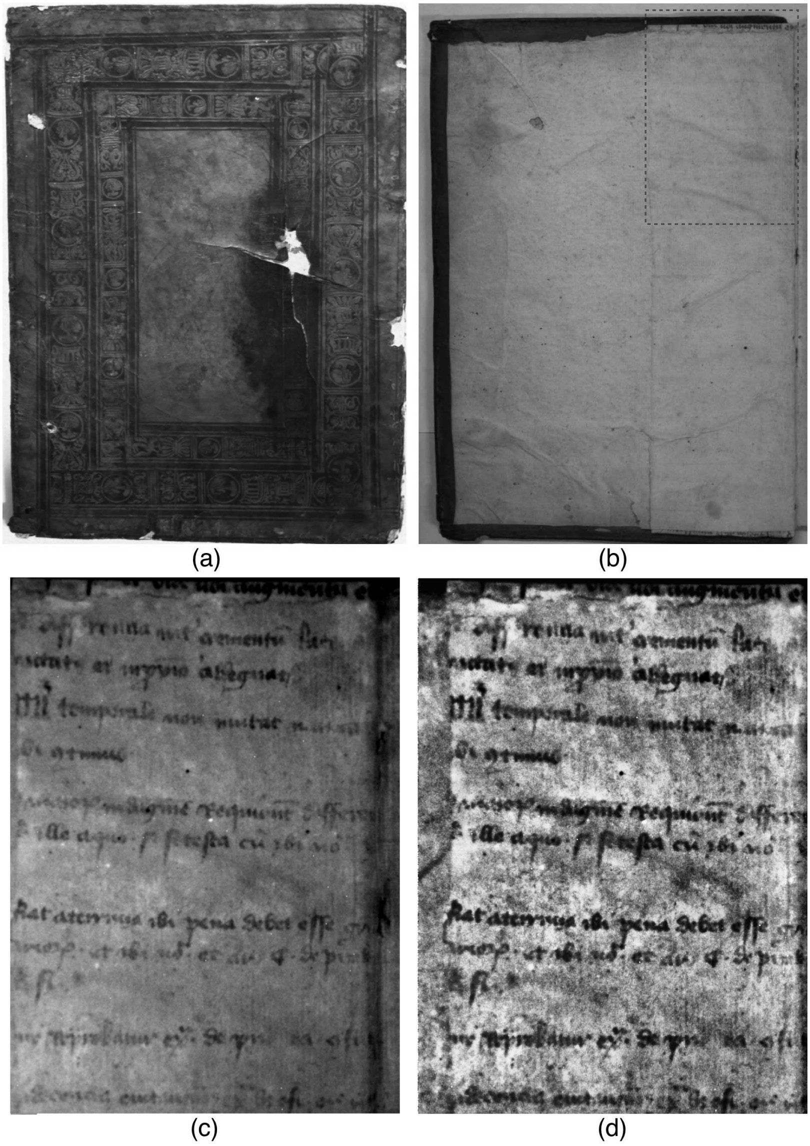


Figure 2. Hyperspectral text recovery of a 16th century book cover showing (a) the exterior, (b) the interior pastedown, (c) recovered text, and (d) thresholded text. Source: Russell Library, Maynooth, Co. Kildare.

13th century copy of the *Etymologies* of Isidore of Seville. The manuscript, although damaged over the centuries, retained most of its original medieval binding structure, and it is reported that part of the project included the compilation of technical evidence that would lead to accurate localisation of the text and a reconstruction of the binding history of the volume. One aspect of this work included the removal of apparently blank pastedown from the upper board; however, disassembly revealed that the pastedown contained writing (cursive Latin text) on the verve. Quandt concludes that manuscript fragments, such as those used as pastedowns, are potentially important, as they may serve to document the origin of the manuscript and its medieval binding.

In order to investigate non-invasive hyperspectral techniques for text recovery, we obtained an exterior 16th century book cover (located in the Russell Library at National University of Ireland Maynooth). This had become degraded with time and suffered from mould in places (Figure 2(a)). The exterior cover, which is intact for the most part, is not of interest here. The interior cover's structure, shown in Figure 2(b), consists of an underlying text which has been pasted over with a clean blank faced sheet of paper. Using fluorescence spectroscopy, i.e, light induced fluorescence in the page, it is possible to reveal the underlying text as shown in Figure 2(c) and 2(d). The underlying text is assumed to be degraded with time but is unreadable due to the presence of the overlaying, pasted down sheet.

When the cover was illuminated with high energy light (505nm) it was found that this light produced fluorescence in the pasted down and underlying pages. Both pages absorb part of this incident light, and in turn emit a low energy (red) light with wavelength 720nm. Using an appropriate filter, our camera records a greyscale image of the resultant fluorescence intensities. As a result of the ink having very low fluorescence relative to the paper, any portion of the interior cover that contains the ink will not fluoresce at 720nm, and is represented by black pixels in our greyscale image, thereby revealing the underlying text.

Following initial light spectroscopy images, we apply hyperspectral analysis techniques and algorithms to further enhance the contrast between page and text and increase the legibility of the recovered text. As described by Salerno et al. techniques have been developed to model the book cover and page as a series of different layers. These layers represent different patterns in the hyperspectral image such as the clean text, the mould pattern and the original parchment or paper pattern. The digital restoration of this book cover involves differentiating the text pattern from interfering patterns and the parchment pattern. This is typically achieved using statistical and image processing algorithms such as Principal Component Analysis (Jolliffe) and Independent Component Analysis (ICA) (Hyvärinen).

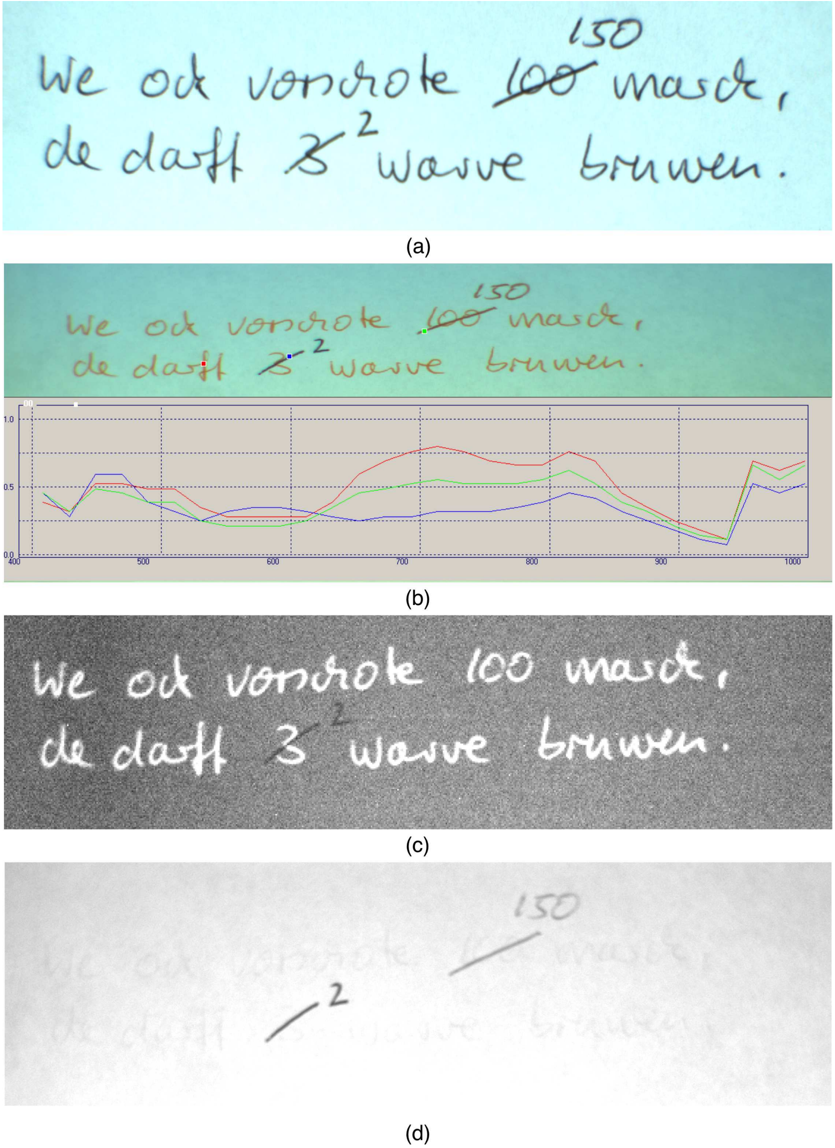
Principal component analysis (PCA), for example, is a statistical analysis technique that is generally used to condense high dimensional data into data of a lower dimension. PCA can also be used to identify the most representative elements of the data, called Principal Components, which is of use in hyperspectral image analysis. These principal components are selected and ranked by their relative variance, judging the most important representative element to be the most variant. Selectively removing certain principal components, and emphasising others, allows the isolation of each of the layers in a hyperspectral image.

In the hyperspectral analysis of historical documents it is seen that individual pixels can be a combination of various substances. In our example, a mould pattern has formed spatially covering different pixels. Pixels spatially covered by this mould pattern will have a spectral signature different to that of pure ink pixels, and pure mould pattern pixels. This type of pixel is termed to be a spectrally mixed pixel. The spectral (pixel) un-mixing of the signals leads to the determining of the contribution of each material in the mix. Un-mixing hyperspectral image data can be seen as an unsupervised method for blind source separation where the objective is to determine the contribution of each component in the mixed signal without prior knowledge of the sub-components. Independent component analysis (ICA), sometimes referred to as *blind-source separation*, is an unsupervised source separation process which provides one method to unmix the different components of a mixed pixel. ICA can be applied to hyperspectral images where the data consists of linearly mixed signals; for this example, we can use ICA to separate the book cover into the aforementioned layers of the cover model. This enhances the readability of the recovered text by separating it from any interference patterns.

In the case of the treatment of the *Etymologies of Isidore of Seville* copy mentioned above, the actual binding structure was not harmed in any way, and the recovered evidence proved to be extremely useful in reconstructing the history of the medieval manuscript. The authors admit that the uncovering and removal of binding fragments seemed at times to be too invasive a procedure. We have shown that the employment of hyperspectral methods for text recovery is a viable option during times when treatment may be too invasive a procedure. The procedures, albeit non-invasive, are complex, non-automated, and require further investigation into appropriate humanities research questions in this domain prior to the production of fitting computer software.

5 Experimental Investigation: Text Segmentation of Different Inks

The general objectives of cursive text segmentation include tasks such as word spotting, text/image alignment, authentication and extraction of specific fields (Likforman-Sulem). An important step associated with all of these tasks is the segmentation of the document into logical units, for instance text lines, words or letters. In general, this is difficult due to the low quality and complexity of these documents, and automatic text



(a)

(b)

(c)

(d)

Figure 3. Hyperspectral segmentation of simulated textual evolution using three different inks; (a) the test sample, (b) hyperspectral spectrogram, (c) original text (white), and (d) later annotations (in black and grey).

segmentation of such kind is an open research field (see also in this volume: Aussems – Brink, Ciula and Stokes). Sophisticated image processing of single-image documents is hence the norm so far (Likforman-Sulem). Here we describe our recent approach towards segmentation of a different kind, which we refer to as hyperspectral segmentation; the technique is based on the separation and segmentation of different inks by recording and analysing their reflectance properties. This technique is particularly useful for the segmentation of texts that have been edited by various authors over a long time period. Thus, it helps in answering basic palaeographic questions and allows the dating of text by comparing the segments with known dates, or using repositories containing hyperspectral properties of different materials (e.g. inks, paper, etc.).

Figure 3 shows a simulation of the textual evolution of *kundige bok* (Rehbein). It is a sentence from this medieval town law, though re-written on modern paper with modern inks for demonstration purposes. As can be seen in the sample (Figure 3(a)), it is hardly possible for the human eye (if at all) to detect whether the changes of the text were made with a different ink or not, thus failing to give an indication whether the two changes within the sentence originate from different points in time in the writing process or not.

In this recently conducted experiment, the hyperspectral scans revealed the different inks surprisingly easily. Figure 3(b) shows the spectrogram of the simulation, measuring the reflectance on three different pixels on the manuscripts, marked green, blue and red. While the spectrograms for two pixels (green and blue) are very close to each other, the third one (red) shows a significant difference. The change, made at the position indicated by the red pixel ('3' substituted by '2') was likely made with a different ink than the change indicated by the blue pixel ('100' substituted by '150'), while the latter was likely made with the same ink with which the original text was written (green dot). Taking into account that medieval scribes produced their ink individually (Wattenbach, 240; Hoheisel 102), using traditional recipes, it can be concluded—if the simulation was based on a real medieval manuscript—that the sample was original text (green) with a simultaneous correction (blue) and a later revision (red). In the *kundige bok* case study we refer to here, this would be an important step towards a complete revelation of the textual evolution and with it the development of medieval town law in late 15th century—information that was not known before.

These preliminary and experimental results thus give hope that work on the original manuscript of *kundige bok* or a similar text, intended to be undertaken in co-operation with An Foras Feasa, National University of Ireland, Maynooth and the Stadtarchiv Göttingen, in Summer 2009, will lead to promising results also. Furthermore, we are also encouraged by the recent results of Klein et al. (2006; 2008) who successfully combine HSI, mathematical feature extraction and classification techniques to analyse documents where several types of ink have been applied, and documents where one ink displayed various degrees of degradation.

Identifying segments of the text written with the same or different ink is only the first step, however, and does not solve the dating issues by itself. It must be accompanied by the expert's view on the manuscript. Dating requires one more piece of information. Consider, for instance, an entry that is undated but from which hyperspectral analysis reveals the same ink signature as a dated entry elsewhere in the book, or even a different source. It can then be dated by inference. However, building up a database of historical ink signatures in a certain (chronological and/or local) context, could establish the basis for an (semi-) automatically created facsimile edition of a manuscript (be it medieval or modern) by visualisation of the different stages (see Figure 3(c)) of the text and could also lead to automated markup preparation—catering, for example, as a tool for the creation of genetic editions.

6 Conclusion

In conclusion, our initial experimental investigations demonstrate the advantages of high-spatial reflectance and fluorescence spectroscopy measurement for the non-invasive examination of historical documents. In particular, it can support codicology research by revealing binding structure of a codex or creating a database of ink signatures, and palaeographic research by making visible hidden text or by giving support to identify scribes and to solve dating issues. We believe that the inclusion of hyperspectral imaging devices as standard research equipment for usable non-destructive analysis of historic documents is both affordable and attainable and would encourage humanities research institutes, libraries and archives to invest in the technologies, methodologies and proficient personnel to maximise their potential. Furthermore, we believe that equitable humanities computing partnerships are an essential component in hyperspectral imaging projects in order to provide realistic *use cases* for the development of the necessary software tools to support disruptive codicology and palaeography research.

Bibliography

Ashutosh, Sharma and Stephen G. Schulman. *An introduction to Fluorescence spectroscopy.* New York: John Wiley & Sons, 1999.

Bansa, Helmut. "Accelerated Aging of Paper: Some Ideas on its Practical Benefit." *Restaurator* 23.2 (2002): 106–117.

Barrow, William J. *Manuscripts and documents. Their deterioration and restoration.* 2. ed. Charlottesville: Univ. Press of Virginia, 1976.

Bischoff, Bernhard. *Paläographie des römischen Altertums und des abendländischen Mittelalters*. Berlin: Schmidt, 1979.

Brannahl, Günther and Malte Grause. "Untersuchungen an Tinten." *Archivalische Zeitschrift* 70 (1974): 79–98.

Brown, K. L. and R. J. H. Clark,. "The Lindisfarne Gospels and two other 8th century Anglo-Saxon/ Insular manuscripts: pigment identification by Raman microscopy." *Journal of Raman Spectroscopy* 35 (2004): 4–12.

Cahill, Thomas. A., Bruce H. Kusko, and Richard N. Schwab. "Analyses of inks and papers in historical documents through external beam PIXE techniques." *Nuclear Instruments and Methods* 181 (1981): 205–208.

Chang, Chein-I. *Hyperspectral Imaging: Techniques for Spectral Detection and Classification.* New York: Kluwer Academic, 2003.

Chang, Chein-I. *Hyperspectral Imaging: Signal Processing Algorithm Design and Analysis.* New York: John Wiley and Sons, 2007.

Clarke, Mark and Marieke Mejers. "Simplification of near-infrared visualisation techniques for identifying blue pigments in situ on manuscripts." *Care and Conservation of Manuscripts* 6 (2000): 242–249.

Erastov, Dimitri. P. "Optico-photographic methods in research in manuscripts and in the visualisation of invisible texts." *Care and Conservation of Manuscripts* 3 (1997): 52–62.

De la Rosa, J. and F. J. Bautista. "Optical properties of paper at 337.1nm." *Revista Mexicana de Física* 51.1 (2005): 110–113.

Fischer, Christian and Ioanna Kakoulli. "Multispectral and hyperspectral imaging technologies in conservation: current research and potential applications." *Reviews in Conservation*, 7 (2006): 3–16.

Fuchs, Robert. "The history of chemical reinforcement of texts in manuscripts – What should we do now?" *Care and Conservation of Manuscripts* 7 (2002): 159–170.

Hapke, Bruce. *Introduction to the Theory of Reflectance and Emittance Spectroscopy.* Cambridge: Cambridge University Press, 2001.

Havermans, John B. G. A., Abdul Aziz, Hadeel, and Scholten, Hans. "Non destructive detection of iron-gall inks by means of multispectral imaging, Part 2: Application on original objects affected with iron-gall-ink corrosion." *Restaurator*, 24 (2003): 88–94.

Hoheisel, Peter: *Die Göttinger Stadtschreiber bis zur Reformation. Einfluß, Sozialprofil, Amtsaufgaben.* Göttingen: Vandenhoeck & Ruprecht (Studien zur Geschichte der Stadt Göttingen, 21), 1998.

Hyvärinen, Aapo, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*, New York: John Wiley and Sons, 2001.

Jolliffe, Ian T. *Principal Component Analysis.* New York: Springer-Verlag, 2002.

Klein, M. E., et al. "The Quantitative Hyperspectral Imager – A Novel Non-destructive

Optical Instrument for monitoring Historic Documents." *International Preservation News* 40 (2006): 4–9. <http://www.art-innovation.nl/fckfiles/file/Downloads/Articles/2006/2006_The_Quantitative_Hyperspectral_Imager.pdf>.

Klein, M. E., et al. "Quantitative Hyperspectral Reflectance Imaging." *Sensors* 8 (2008): 5576–5618. <http://www.art-innovation.nl/fckfiles/file/Downloads/Articles/2008/sensors-38-02-original.pdf>.

Likforman-Sulem, Laurence, Aderrazak Zahour, and Bruno Taconet. "Text line segmentation of historical documents: a survey." *International Journal on Document Analysis and Recognition* 9.2-4 (2006): 123–138.

Missori, M., M. Righini, and S. Selci. "Optical reflectance spectroscopy of ancient papers with discoloration or foxing." *Optics Communications* 231 (2004): 99–106.

Mitchell, Charles Ainsworth. "Inks. Their composition and manufacture, including methods of examination and a full list of British patents." 4. ed. London: Griffin & Co, 1937.

The National Academy of Sciences. *Preservation of Historical Records*, Chapter 4: "Paper", 1986. <http://www.nap.edu/openbook/030903681X/html/33.html>.

Ó Cróinín, Dáibhí, ed. *A New History of Ireland: Prehistoric and early Ireland*, Oxford: Oxford University Press, 2005.

Padoan, R., et al. "Quantitative Hyperspectral Imaging of Historical Documents: Technique and Application." *ART Proceedings* (2008). <http://www.art-innovation.nl/fckfiles/file/Downloads/Articles/2008/padoan2008.pdf>.

Penders, Nathalie J. M. C. and John B. G. A. Havermans. "Preventive conservation related to iron-gall ink deterioration." *Care and Conservation of Manuscripts.* 6 (2000): 18–32.

Quandt, Abigail B. "The Documentation and Treatment of a Late 13th century Copy of Isidore of Seville's Etymologies" *The Book and Paper Group Annual.* Ed. Book and Paper Group (BPG) of the American Institute for Conservation of Historic and Artistic Works, Volume 10, Washington, D.C., 1991. <http://aic.stanford.edu/sg/bpg/annual/v10/bp10-15.html>.

Rehbein, Malte. "Reconstruction the Textual Evolution of A Medieval Manuscript." *TEI@20. Special Issue. LLC. The Journal of Digital Scholarship in the humanities.* (forthcoming).

Roselieb, Hans. "Die Chemie alter und neuer Tinten." *Archivalische Zeitschrift* 70 (1974): 74–78.

Salerno, Emanuele, Anna Tonazzini, and Luigi Bedini. "Digital image analysis to enhance underwritten text in the archimedes palimpsest." *International Journal on Document Analysis and Recognition* 9.2–4 (2007): 78–87.

Scholten, J. H., M. E. Klein, and Th. A. G. Steemers. "Hyperspectral Imaging – Concepts and Potential in Paper and Writing Durability Research." *Proceedings Durability of*

Paper and Writing. Ljubljana, 2004. <http://www.art-innovation.nl/fckfiles/file/Downloads/Articles/2004/2004_Hyperspectral_Imaging.pdf>.

Scholten, J. H., M. E. Klein, Th. A. G. Steemers, and G. de Bruin. "Hyperspectral imaging - a novel nondestructive analytical tool in paper and writing durability research." *Proceedings Art 05*. Lecce, 2005. <http://www.art-innovation.nl/fckfiles/file/Downloads/Articles/2005/2005_Hyperspectral_Imaging.pdf>.

Smith, K. J. *Evaluation of Whiteness and Yellowness in Color Physics for Industry*. 2nd Edition, Society of Dyers and Colourists, 1997.

Tilley, Richard. *Colour and the Optical Properties of Materials: An Exploration of the Relationship Between Light, the Optical Properties of Materials and Colour*. New York: John Wiley & Sons, 2000.

Vaarasalo, J. "Optical Properties of Paper in Papermaking Science and Technology." 17. *Pulp and Paper testing*. Eds. J. Gullichsen and H. Paulapuro. Helsinki: Fapet Oy, 1999: 162–181.

Van der Reyden, Dianne. "Recent Scientific Research in Paper Conservation." *The Journal of the American Institute for Conservation* 31 (1992): 117–138. <http://www.si.edu/MCI/downloads/RELACT/paper_properties_degrade.pdf>.

Wattenbach, Wilhelm. *Das Schriftwesen im Mittelalter*. Zweite, verm. Auflage. Leipzig: Hirzel, 1875.

Wunderlich, Christian-Heinrich. "Geschichte und Chemie der Eisengallustinte. Rezepte, Reaktionen und Schadwirkung." *Restauro* 100 (1994): 414–421.

Zerdoun Bat-Yehouda, Monique. *Les encres noires au Moyen Âge. Jusqu'á 1600*. Paris: Éd. du Centre Nat. de la Recherche Scientifique (Documents, études et répertoires publ. par l'Institut de recherche et d'histoire des textes), 1983.

Aspects of Application of Neural Recognition to Digital Editions

Daniele Fusi

Abstract

Artificial neuronal networks (ANN) are widely used in software systems which require solutions to problems without a traditional algorithmic approach, like in character recognition: ANN learn by example, so that they require a consistent and well-chosen set of samples to be trained to recognize their patterns. The network is taught to react with high activity in some of its output neurons whenever an input sample belonging to a specified class (e.g. a letter shape) is presented, and has the ability to assess the similarity of samples never encountered before by any of these models. Typical OCR applications thus require a significant amount of preprocessing for such samples, like resizing images and removing all the "noise" data, letting the letter contours emerge clearly from the background. Furthermore, usually a huge number of samples is required to effectively train a network to recognize a character against all the others. This may represent an issue for palaeographical applications because of the relatively low quantity and high complexity of digital samples available, and poses even more problems when our aim is detecting subtle differences (e.g. the special shape of a specific letter from a well-defined period and *scriptorium*). It would be probably wiser for scholars to define some guidelines for extracting from samples the features defined as most relevant according to their purposes, and let the network deal with just a subset of the overwhelming amount of detailed nuances available. ANN are no magic, and it is always the careful judgement of scholars to provide a theoretical foundation for any computer-based tool they might want to use to help them solve their problems: we can easily illustrate this point with samples drawn from any other application of IT to humanities. Just as we can expect no magic in detecting alliterations in a text if we simply feed a system with a collection of letters, we can no more claim that a neural recognition system might be able to perform well with a relatively small sample where each shape is fed as it is, without instructing the system about the features scholars define as relevant. Even before ANN implementations, it is exactly this theoretical background which must be put to the test when planning such systems.

Zusammenfassung

Künstliche neuronale Netze (Artificial Neural Networks, ANN) sind in solchen Softwaresystemen weit verbreitet, die Probleme wie Zeichenerkennung zu lösen suchen,

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 175–195.

ohne einen traditionellen algorithmischen Ansatz zu verfolgen. Weil ANN am Beispiel lernen, brauchen sie einen konsistenten und gut ausgewählten Satz an Proben um die Mustererkennung zu trainieren. Das Netz reagiert mit einer hohen Aktivität in seinen Output-Neuronen, wenn ihm ein Input-Muster präsentiert wird, das zu einer bestimmten Klasse (z.B. einer Zeichenform) gehört. Es kann die Ähnlichkeit von Beispielen berechnen, die mit einem solchen Modell noch nicht verglichen worden sind. Typische OCR-Anwendungen erfordern eine erhebliche Vorverarbeitung der Beispiele, bei der die Größen der Abbildungen verändert und das »Rauschen« (d.h. störende Daten) entfernt werden, so dass sich die Buchstabenformen klar vom Hintergrund abheben. Darüber hinaus braucht man normalerweise eine sehr große Zahl von Beispielen, um das Netz darauf zu trainieren, ein Zeichen von allen anderen zu unterscheiden. Dies stellt wegen der relativ niedrigen Qualität und der hohen Komplexität der verfügbaren digitalen Beispiele eine Herausforderung für paläographische Anwendungen dar. Es wirft noch mehr Probleme auf, wenn es darum geht, feine Unterschiede, z.B. die besondere Form eines einzelnen Buchstabens aus einer bestimmten Phase eines Scriptoriums, zu entdecken. Dabei scheint es ratsam, einige Regeln zu bestimmen, nach denen die für bestimmte Fragestellungen wichtigsten Merkmale aus den Proben extrahiert würden, und das Netz dann nur noch mit einer Teilmenge der andernfalls überwältigenden Menge an feinen Unterschieden arbeiten zu lassen. ANN sind keine Zauberei, und es bedarf immer des sorgfältigen Urteils der Forschenden, um eine theoretische Grundlage für ein computergestütztes Werkzeug zu schaffen, das bei der Lösung ihrer Probleme helfen soll: Dies lässt sich leicht an anderen Beispielen zu IT-Anwendungen in den Geistes-und Kulturwissenschaften belegen. So wie wir kein Hexenwerk bei der Erkennung von Alliterationen erwarten können, wenn wir ein System nur mit einer Reihe von Buchstaben füttern, so können wir auch von keinem neuronalen Erkennungssystem verlangen, dass es mit einem relativ kleinen Satz an Beispielen zurecht kommt, wenn jedes Zeichen unbearbeitet eingegeben wird, ohne das System darüber zu informieren, welche Merkmale von den Forschenden als besonders wichtig definiert werden. Ein solches theoretisches Fundament muss jedoch noch vor der Planung und Umsetzung künstlicher neuronaler Netzwerke entwickelt und geprüft werden.

1 Scenario

This paper derives from a much wider discussion of a digital edition system.[1] I have been creating the collection of a large amount of different data related to a specific subject. One of the chief points I often stress about the principles of this system is that its purpose is to create a “truly” digital edition of textual and/or non-textual material,

[1] For what follows see e.g. Fusi 2007 (principles for the epigraphical system) and Fusi 2008 (expert metrical system), together with their hosting website.

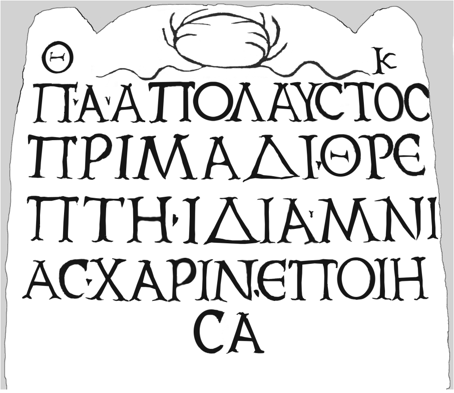


Figure 1. Digital Drawing.

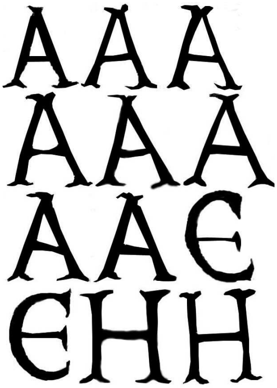


Figure 2. Some shapes extracted from the drawing.

which should also be treated as a research tool rather than as a mere clone of a traditional paper edition. Also, the system aims to be greatly expandable so that new, often very specialized, data can be added to the existing material at any time, attaching to its structure rather than modifying it. For instance, a collection of epigraphical texts can be enriched by specialized data about linguistics, metrics, history, prosopography, archaeology, palaeography, etc. which may eventually come from expert systems capable of providing automated analysis (like in the case of metrical analysis); also, some of the existing data, e.g. a collection of digital drawings taken from photographs, can also lay out the foundation for further developments which may sound particularly interesting in fields like palaeography. Here, I would just like to discuss some basic aspects of one of these suggested applications, exploring the feasibility of applying neural techniques to graphical data: among others, a further application for the multimedia capabilities of digital editions can be provided by pattern recognition using neural methods. Typically we can start with a sample of inscriptions, even if such methods can be applied to several other scenarios (e.g. manuscripts, brick stamps, iconography, coins, etc.), probably with different levels of efficiency and different requirements for data preparation. In the hypothetical scenario of an epigraphical edition already provided with several digital drawings extracted from their photographs, the next step might be to extract the shapes of each letter from each of them, and then feed a neural recognition system with them as samples for a given period or region.

For instance, we could extract all the letter forms from selected Greek inscriptions in our corpus (see the Figures 1 and 2) and use them to train a neural system, providing it with different chronological classes. Once this is done, we could imagine a publishing scenario where users perform a truly graphical query to retrieve all the inscriptions whose letters resemble a provided sample, which might come from a photo, or from a drawing sketched by the user himself with an inking system; such a system might then provide a first indication for the dating of the inscription the letter sample has been taken from, or just any other type of palaeographical classification we have defined as relevant in our texts. A by-product of this system (or rather a prerequisite for it) might also be a detailed catalogue listing all the shapes of all the palaeographically relevant inscriptions in a corpus, where users can browse them on screen century by century, or according to any other filtering or sorting criterion, and immediately get a feeling for the evolution of the writing system. This is just a small example but the possibilities might be endless; here I would like to stress the importance of fully exploiting the abilities of a true digital edition. Automatic pattern recognition is just another sample of these abilities.

2 Pattern Recognition and ANN

In today's software, pattern recognition (including, of course, optical character recognition, OCR) is typically accomplished using artificial neural networks (ANN). ANNs are a complex subject, and even an introduction to them would be beyond the scope of this work, therefore I'll sketch out some of their principles with reference to their practical application to software systems[2]. The easiest method to grasp the way an ANN works is to compare its behaviour to traditional software solutions which typically use an algorithmic approach for solving problems: there is a sequence of instructions to follow, and this, of course, implies that we must know them in advance. Instead, neural networks, composed by interconnected elements (neurons) working in parallel, learn by example, so that they cannot be programmed to perform a specific task. This implies that selecting the right examples and defining the relevance of their traits is a crucial point.

An artificial neural network is a set of nodes and connections between them; the nodes (neurons) are the computational units: they get some input and process them to produce an output. It is the right interaction of nodes through their connections which leads to defining an emergent behaviour for the network, so that its abilities supersede those of its elements. These artificial neurons are inspired by natural neurons, which receive signals through synapses; when the strength of the signals exceeds a certain threshold the neuron is activated and emits a signal through the axon. This signal in turn might be sent to another synapse, and might activate other neurons, and so

[2] Introductions to ANN abound in the web. Here I'll mainly follow (with consistent simplifications) Stergiou and Siganos, and Gershenson. Other material can be found in the documentation of the API of several open-source and/or freeware software libraries implementing ANN or their derivatives (see e.g. Chang and Lin LIBSVM website for a list of ports of this software library in several languages).

on. An artificial neuron is a device with several inputs and a single output. These inputs are multiplied by weights (strength of the respective signals) and then computed by mathematical functions which determine if a neuron should be activated or not. Another function, which may or not be different, calculates the output of the neuron, often related to a specific threshold. By adjusting the weights of a neuron we can obtain the output we desire for each specific input. As a typical ANN contains many neurons, it would be very complex to do such adjustments by hand, so special algorithms are used; this process of adjusting weights represents the training or learning of the ANN. A neuron thus operates in training mode or in usage mode: in the training mode it is trained to fire, or not, for particular input patterns, i.e. to associate a specific input to a specific output. In usage mode, a neuron either encounters a taught pattern or a new one; in the former case it simply fires the taught output, in the latter case a firing rule is used to determine whether to fire or not.

In connection to our subject it is to be stressed that a firing rule relates not only to the taught input patterns, but to all the input patterns. For instance, a firing rule can be implemented with a technique which compares elements in patterns so that the neuron fires or not according to whether the compared patterns have more input elements in common with the nearest pattern in the firing-taught set, or in the not-firing-taught set. Such a rule provides the neuron with the "sense" of similarity so that it can respond to patterns never encountered before, returning the output corresponding to the taught input pattern which is most similar to the given pattern. Also, more complicated neurons can be used, where inputs are "weighted" by multiplying the input values of a pattern by a specific number and adding their results, letting the neuron fire only when the sum exceeds a predefined threshold value. Such neurons can adapt to specific situations by changing their weights or threshold; there are several algorithms for adapting, and one of the commonest is the back error propagation, where we start with randomly chosen weights and then adjust them so that the error is minimal.

A very common type of ANN is built of three sets of neurons: a layer of input neurons is connected to a middle layer of "hidden" neurons, which in turn is connected to a layer of output neurons. The activity in the input layer represents the input data fed into the network; the activity of the middle layer is defined by the activity of the input layer and the weights assigned to the connections between the input and the middle layer; and in turn the activity of the output layer is defined by the activity of the middle layer and the weights assigned to the connections between the middle and the output layer. In this model the middle layer is thus free to build its own representation of the input, by connecting one or more neurons of the input layer to one or more neurons of the middle layer and assigning various weights to these connections. The input pattern units can thus be variously grouped, combined, selected and weighted according to the specific problem the ANN should solve.

These connections and their weights define the “knowledge” of an ANN network; modifying this knowledge according to experience implies a learning rule which can change these weights. Typically for pattern recognition the learning is supervised, in the sense that it happens by associating input samples (e.g. the various shapes of the character A) with their desired output (the character identified as A). For instance, say we want to recognize the 26 printed lowercase letters of the standard ASCII alphabet (a–z): we might use a grid of 16×16 optical sensors, each capable of detecting the presence, or absence, of ink on a small portion of the printed character area. We would thus need an input layer of 256 units (16×16) and an output layer of 26 units (one for each letter): for each letter the network should produce high activity in the corresponding output neuron and low activities in all the other output neurons. For instance, for a correctly recognized letter c there might also be a fair amount of activity in the output neuron corresponding to the letter o, which “looks” like c, but the activity in the output neuron corresponding to c should be considerably higher. To train the network we would present it a set of sample images for each letter, compare the activity produced in the output layer by each of them and calculate the error (defined as the square of the difference between the actual and the desired activities). We would then use some algorithm to adjust the weights of each connection in order to reduce the error to its minimum and thus get the best recognition results.

As shown in this sample, we are thus representing each letter shape with a set of units, i.e. a vector of numerical values: just think of superposing an ideal grid to a printed character, and mapping each grid cell to a number representing the presence (e.g. 1), or absence (e.g. 0), of ink on a small portion of the printed character area. For instance, if we were using a 3x5 grid we might represent a letter A like: in Figure 3, which in turn might be represented by a vector of numbers like [0,1,0, 1,0,1, 1,1,1, 1,0,1, 1,0,1]. Of course, in a generic OCR scenario a first issue is represented by the amount of preprocessing required to extract such a black and white silhouette from a photographic image: first of all, we must reduce the amount of non-essential information, which for a digital image typically means resizing each character image to a predefined size using the proper resampling algorithm, removing colour information, reducing a greyscale image to a black and white one and applying any adjustment technique considered useful for letting the image contours emerge clearly from the background “noise”. This step can be very complex and implies a lot of digital image processing techniques. Also, the size of the character should be accurately chosen so that it is not too small (otherwise complex shapes might result into a barely readable black spot once resized) nor too big (which might result in vectors too big to be manipulated in learning and recognition). Nevertheless, it might also happen that our chosen vector size still impacts the performance of our network:

	X	
X		X
X	X	X
X		X
X		X

Figure 3.

indeed, it is easy for a vector size to grow when representing letters, especially when their nuances are essential. Even a very small grid like 16×16, which would probably be too small to be usable for palaeographical purposes, implies 256-units vectors, and just doubling the grid would already take us to 1024 units. Computers are fast, but if you think of the complexity and number of calculations implied by the usage of an ANN built of at least 3 interconnected layers with several input characters each represented by thousands of units, which must be trained by adjusting the weights of the whole system even hundreds of times, it is easy to understand that we might incur very serious performance problems. To this end, OCR specialists often use additional processing techniques like receptors: for instance, think of a grid of points representing a character shape; instead of using vectors containing a unit for each point we might imaginarily draw a set of line segments over the shape, with arbitrary size and directions, and state that a receptor has an activated value when it crosses a letter and a deactivated value when this does not happen. This way we would just have as many units per vector as receptors (i.e. lines). This technique (which, of course, implies the problem of generating such receptors in an efficient way) may be very powerful in some contexts where we can be satisfied in detecting the essential traits of each character in order to distinguish them, but it does not fit well where our shapes are rather complex and our objective is to pay attention to finer nuances.

For such applications it is probably better to consider a relatively recent outgrowth of ANN, support vector machines (SVM), which are close to ANN but typically outperform them when dealing with specific problems like pattern recognition; SVM work by finding the optimal "boundary" ("hyperplane") which divides groups of vectors (i.e. the set of features which define a pattern) so that (all or most of) the ones belonging to a recognized class (e.g. a letter, when recognizing characters) stand on one side of the plane, while all the others stand on the other side (the vectors along this "boundary" are named support vectors, whence the term). As for ANN, there are several open-source libraries available for building applications using such techniques; a sample of a trivial application based on one of these libraries (built in C# and just slightly modified by myself for easier use in an asynchronous environment) is represented by the example shown here.

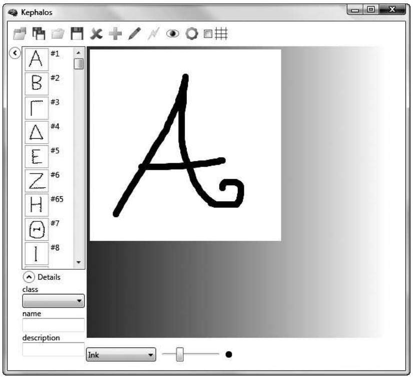


Figure 4.

In this demo application, users literally draw a set of shapes (which might represent letters or any other thing), assign them a class (e.g. tell the software that the given sample represents an A, another a B, etc.), and let the system learn from them by getting its own "idea" of each of the declared classes. Once the system is trained, the user can draw a new shape and ask the system to identify it.

3 Application Issues

The above sample is just a demonstration, but by its very nature, this is a field where it is very difficult to make previsions on the efficiency of any given real-world system before actually putting it to the test. From a practical standpoint, detecting patterns using such methods poses several problems. First of all, as we have seen, all these neural approaches require very careful training and consequently a very high number of samples (i.e. images, in our case): these techniques solve problems by learning to associate (in various complex ways) input patterns (e.g. a given shape for letter A) with output patterns (the letter A in itself): there is no algorithmic approach here, as (fortunately!) we are not going to tell the machine which all the steps are to recognize each shape as a specific letter. We just "show" it samples for each letter, and let it "figure out" to which of these samples any given sample is the nearest. This is, of course, the only way of handling such problems rather than thinking exclusively in algorithmic terms, which might easily turn the desired solution into a programmer's nightmare, even supposing that we can imagine a set of well-defined and ordered steps to instruct a machine to distinguish each single shape from all the others (and what if we decide to add new classes to be recognized?). Usually such techniques are used for experimental data which can be increased at will, and often happen to be so overwhelming in number that manual processing would be impossible. This may represent an issue for palaeographical or other similar applications because of the relatively low quantity of digital samples available, and poses even more problems when our aim is detecting subtle differences rather than essential patterns. For instance, if we want to train a system so that it recognizes the differences between the same letter *alpha* with or without serifs, or with a broken rather than continuous horizontal segment, rather than training it mainly for distinguishing between different letters (an *alpha* is more obviously different from an *omicron*), this would require even more samples, and might possibly still result in an excessive number of system failures. So, for real-world applications of such techniques to palaeographical data we must take into account several potential issues:

1. The preprocessing required to obtain vectors describing each shape we want to use for training or recognition from a photographic image (resizing, dealing with colour or greyscale information, detecting edges, removing noise, etc.) is very complex. Also, the required graphical preprocessing especially for manuscripts should cope with a very "noisy" background and with very different sizes, shapes,

and locations of letters. In some scenarios it might be possible to exploit existing digital data if available, but even then the image-letter(s) pairs to be used to feed the training process should meet some basic requirements (in terms of resolution, colours, shape isolation, etc.) which would probably require additional processing.

2. The scenario is even worse when applied to manuscripts rather than to epigraphical material, as for obvious reasons in the former case the shapes are usually much more complex and ligatures and abbreviations define new graphical signs which we should treat like individual letters, as they may bear no resemblance to their simple components. Thus, the target alphabet to recognize would be much bigger, including not only the simple letters but also several of their peculiar combinations; in other terms, we would have a much bigger output layer in our network.
3. Also, if we want to experiment with different variants of the same shape, things get even worse as the nuances in the general context of a shape may be so subtle that this would require a very troublesome learning process, the required samples would probably exceed the size of the samples we could provide, the processing performance would be seriously degraded and the results inadequate, as there would be too much noise activity in the output layer. The very nature of such techniques relies completely on learning by samples, and when samples belonging to different classes are too similar each other, it becomes impossible for the machine to grasp their differences; we would have to raise the similarity thresholds up to a point where each sample would be judged as representative of a different class. In this case, the only approach should probably be a much wiser preprocessing, even manually driven, which selects just the traits we judge as relevant for our purposes: but a similar approach might easily become too expensive to be rewarding.

In a similar scenario, a more proficient approach would take into account a number of helper methods and start from the assumption that neural recognition is best fit to figure out and compare essential patterns from large sample data rather than distinguishing subtle nuances among similar types. In practical terms this means that first of all we would have to think about a number of pre-processing steps which help define a priori which traits of each sample should be treated as more relevant, thus reducing all the noise data present in the original sample. This does not only include generic digital image processing as explained above; on a more specific ground it would be wise to preprocess the resulting images so that the system gets some well-directed clues about the features which scholars themselves define as relevant for the definition of a pattern.

This often sounds odd to people in connection with neural systems, but especially in these contexts such systems probably would never be proficient unless directed, and it does not need to be remarked that any digital tool is not meant to replace scholars but only to be useful for their research. As for any other technology, such systems must be applied to the right problems and satisfy a number of requirements. In a palaeographical scenario it would be necessary to isolate the traits which scholars define as essential for the classification of a sign, dropping all the other features which just produce noise; for instance, this might mean erasing some portions of each sample letter to keep just the traits which are essential to the definition of each sign, and then let the system train on such "mutilated" samples rather than on the full picture. Automatic or semi-automatic specialized systems might be devised for such preprocessing, but the main point is that it is up to the scholar's judgement to decide the kind of material and the edition purposes. In this context, a neural system might prove useful anyway as it provides a means of defining patterns in a more abstract way, without having to explicitly define a fully exhaustive list of models and devise some (maybe complex) digital format to store and compare them. For example, think of a number of letter shapes: if we were just interested in some very selected aspects we could even dispose of images, and rather describe them with a variable set of attributes, which may or not be present in this essential description of each letter (e.g. serifs, horizontal angular traits, shading, etc.), and still have a searchable system without even requiring neural techniques. We might need to go further and take into account a greater number of traits for each shape, and we might not even be able to list them in advance, or to say whether some of them can be treated as equivalent or not, because we still do not know the whole corpus with all its thousands samples. In such cases, it might be useful to provide a neural system with some well-preprocessed samples and let it build some patterns from them, thus providing some sort of fuzzy comparison less constrained to predefined sets of attributes. Of course, the main job of the scholar in this context is defining a theoretical framework which provides the clues to the system about what is more or less relevant in any given type of sample data; this framework might often be just a working hypothesis to start with, as this analysis process typically implies "on the spot" adjustment of a number of sensitive parameters which affect the general appreciation of observed data and repeating the process several times until the best result is attained. This is the typical way of working in most fields where information technology is applied to human sciences: in my opinion, one of the most difficult yet intriguing aspects of such applications is the requirement of a degree of formal definition of every theoretical aspect strict enough to be applicable to machine analysis. Of course, often it is not possible to attain such definitions in theory, but at any rate we must provide at least a good working hypothesis which fits the purpose of our research and works as well as we can reasonably expect in our digital product.

4 Theoretical Frameworks and Preprocessing: A Textual Analogy

The examples could be easily multiplied, but I think it might be useful to quote just one to emphasize this essential point. I have already stressed the concept of a digital "edition" which can be the basis of any other specialized application, like e.g. metrical analysis of the texts collected in some corpus. A component of my expert system devoted to such problems refers to the observation of alliteration, which is much more relevant in Latin than in Greek texts, but can also be used in connection with other languages and researches.[3] Of course, a full discussion of this component would be beyond the scope of this paper, but I would just like to stress the same methodological implications already referred to the problem of applying neural recognition systems to graphical samples.

As alliteration refers to the repetition of sounds,[4] a first naive approach to detect this phenomenon might be counting all the equal sounds in a given context.[5] Even if (for economical reasons) we do not discuss here the issues connected to defining which (even approximate) "sounds" correspond to a sequence of letters, and which of them may be treated as "equal" for the purpose of alliteration, it is obvious that such issues represent the first step in building a theoretical framework for our analysis. We must go further: the phenomenon of alliteration cannot be oversimplified to the point of simply counting the equal sounds in some text. Otherwise, we might end up concluding (with Evans) that hexameter parts like *Pergama Graiis* or *talia fatur* are alliterations just because in the first sample the syllables in strong position share the "letters" *g* and *r*, and in the second they share *t* and *a*, even if their order is reversed, as we are just "counting letters". On these premises, it is hardly surprising that Evans claims that "the lines which satisfy the fundamental rule [...] are very numerous" and that for instance in Ovid's *Fasti* they are "about eighty percent of the whole" (44). Evans himself goes further, noticing that at any rate "owing to the fact that there are only sixteen consonant or vowel sounds which cannot echo each other, it is difficult to construct a long line without a single rhyme" (5). In more modern terms, given that language is articulated, it will be obvious that any text, whatever its extent, will be built with a very limited set of phonemes: It is right the finite and very small number of phonemes which grants language its economy, and it is a trivial prediction that any text will show the repetition of these phonemes. Even to the ears of a more naive listener it should be clear enough that lines like *iam licet venias marite* (Catull. 61, 187) could hardly

[3] The component itself has no dependency on a specific language (nor digital format), even if, of course, the theory behind it fits some well-defined principles; for instance, the component has been used in connection with Latin, Greek and Italian, for both metrical and non-metrical texts.

[4] Even if a true theory of alliteration seems missing from ancient authors, there are some relevant traces of their appreciation of the phenomenon; one of them is the synthetic yet significant definition in *rhet. ad Herennium* 4,2,18 "*nimia assiduitas eiusdem litterae*").

[5] That this approach has not been judged as naive as it might sound is shown by the fact that scholars like Evans did really base a full "theory" of alliteration on it. In his view alliteration is just a subset of what he calls "rhyme", defined as (p. xiii) "an agreement in sound between two or more syllables" which "may extend either to one letter or more", so that even words like *like* and *roll* would "rhyme" because of the shared *l*.

be said to have the same acoustical effect as lines like *tympana tenta tonant palmis et cymbala / circum concava* (Lucr. 2, 618–9). This is, of course, the effect of the fact that the phenomenon cannot be described on a purely quantitative basis (the count of "letters"), as what is relevant is the relative position and distance of the repeated sounds and their distribution into linguistic units (words and syllables), rather than the (trivial) fact that sooner or later they must repeat in a text.

Should we limit ourselves to count the occurrences of each "letter" (or better "sound") in a line, we might grasp the existence of some alliteration when this is rather emphasized, like in the Lucretius line quoted above, but even there it appears that several letter counts just add noise to our picture rather than clearly presenting it, as we can see from the following chart representing the raw letter counts (as percent) in this line:

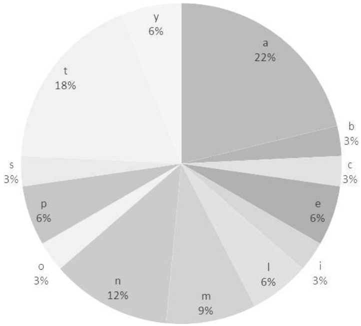


Figure 5.

As you can easily see, the /t/ in this chart does not stand out as it does when we hear this line (even if uttered with sounds which do not correspond to the ancient ones): this happens because the chart assigns the same value to each letter rather than reflecting the hierarchy of the linguistic elements involved, their relations, and the literary tradition at the base of the texts examined. Of all the letters represented in this chart, only very few are really relevant for our purpose, not because of their intrinsic phonetic value but just because of their position in the text examined. Thus, most of the slices in this chart are just noise which distract attention from the phenomenon we have selected for analysis. If we want to provide some score of the level of alliteration detectable in a text, we must first provide at least a theoretical definition of the phenomenon as a working hypothesis, thus ruling out all the irrelevant "letters", just like in ANN each neuron can "weight" its inputs to produce the expected output. To this end, we could broadly state some chief starting points:[6]

[6] For the alliteration proper these points mainly reflect to some extent Valesio; some good examples (col-

Phonemic analysis: alliteration is the repetition of sounds, and as such any text requires analysis in phonemic terms. Of course, this implies an approximate reconstruction of phonemic values for the texts being examined, in this sample Greek and Latin texts, whose limits in terms of chronological and geographical extent are well known, but this is a necessary limitation. This also enforces the assumption that phonetic variants of the same phoneme are treated as alliteration: thus the ground for this analysis is defined in phonological (rather than phonetic) terms. Also, it must be remembered that the Greek and Latin texts analysed were read aloud for an audience rather than silently read, so that we must resist the temptation of analysing this phonetic phenomenon as a printed text, referring to the eye rather than to the ear.

Alliteration and language constraints: it is very difficult to estimate language constraints on alliteration, but in this working model I stress the (primarily word-) **beginning** sounds rather than the ending sounds, for both practical and theoretical reasons. First, cases of *homeoteleuton* (which is often also an *homoioptoton*) are very frequent among connected words in inflectional languages like Latin or Greek (and in general in Indo-European languages) where the grammatical role of suffixes is generally much wider than that of prefixes: the frequency of such “grammatical rhymes” was already noticed and discarded as irrelevant since Evans.[7] Second, the alliteration of word-initial sounds is well-attested as a traditional feature which might well go back to what comparatists call the Indo-European metrics (it is the main metrical feature of systems like the ancient German poetry,[8] and its importance in the archaic Latin poetry is well-known). Also, some morphosyntactic diachronic phenomena clearly show the importance of word-initial sounds in connecting words into combinations which later can be frozen (e.g. Italian *domineddio / domeneddio* from *domine Deus*, or the strong connection between ἦμος and the expression of solar chronology in Homer and Hesiod, enforced by the combination with ἠώς and ἠέλιος[9]).

Of course, for the general sound of the whole verse or sentence also inner- and final (even if much more forced by the language structure) repetitions count,[10] nor is

lected with a fine stylistic sense but in the context of a much weaker theory) are also found in Herescu, who anyway devotes only a section to the alliteration.

[7] “Such rhymes in Latin are merely accidental” (Evans 2, quoting Hor. *ars* 100–101 ending with two imperatives: … *sunto /… agunto*).

[8] Cf. Valesio 25 ff.

[9] Cf. Valesio 179–181 who refers to Monteil.

[10] Valesio, who also restricts (for the same practical reason of economy) his definition of alliteration, really takes into account only sequences of initial or final sounds, regarding every combination of them (i.e. all the sounds must be initial or all the sounds must be final: no combination is taken into account) or any internal repetition at most as “para-allitterazioni”. Instead, he widens his definition when dealing with vowels, thus being forced to abandon the consideration of their relative position, probably also because of the English texts he uses as samples (for some considerations about vowels in Latin poetry see Herescu 84ff., whose approach anyway seems somewhat questionable because of the more limited role of vocalic

discussing the intentions of the poet relevant here,[11] as I am looking for an essential pattern detection function. Also, I prefer to apply to the tradition of the rhyme and its ancestors a different set of analysis functions, which in Classical times play little or no role in Latin and Greek poetry and anyway mostly relates to another context (typically alliteration is an intrastichic and rhyme an extrastichic feature). Thus I mainly define (detect) alliteration in terms of beginning sounds, looking to the other repetitions only at a second stage, and descending from "words" to syllables and phonemes. The main pillars for building an alliterating sequence are thus considered the "words"; inner (syllabic and phonemic) alliterations are taken into account only in dependency of word alliterations. Finally, more generic euphonic considerations are not of concern when dealing with this strict working definition of alliteration: thus I exclude the analysis of repetitions of vowels whatever their position may be, unless they are involved in the alliteration as already defined (e.g. words beginning with the same vowel, syllables beginning with the same vowel, vowels echoing the main alliterating vowel).

- The resulting **hierarchy** of linguistic units for this analysis is as follows: sentence, word, syllable, phoneme. The sentence (or line in metrical texts) defines the analysis context; words beginning with the same sound(s) define alliterating sequences; in each of these sequences I extend the analysis of repeated initial sounds to each syllable of each word, and finally add some weight to the "echoes" of the main sound in the remaining (i.e. not included in initial repetitions) phonemes of each syllable.
- The **context** for alliteration is defined as the single verse in metrical texts and the single sentence in non-metrical texts. This does not rule out the fact that sometimes alliterations can be carried over more lines, but this is a secondary feature when dealing with an essential detection function, and belongs to a wider analysis context which is not my primary concern at this stage.
- **Leftwards scan**: each word in a sequence is compared with any of the words at its left in the same sequence, taking into adequate account the number of the initial segments which are considered equal for the purpose of alliteration and their relative distance. This is compliant with the assumption that most of the ancient poetry is defined in acoustic rather than in visual terms, as it is orally performed for an audience: thus the repetition of sounds becomes apparent once the same sound pattern is repeated by newly uttered words: the word just uttered can be traced

phonemes in the Latin language, but at least takes into account the necessary distinction between "strong" and "weak" syllables – p.26).

[11] On this questionable aspect cf. e.g. Herescu 128: "même si les rencontres phoniques se produisaient d'elles-mêmes, à l'insu de l'écrivain, il reste néanmoins que celui-ci part avec, dans l'oreille, un certain enchaînement de sons, une certaine musique, qui se réalisera dans ses vers même sans qu'il s'en rende compte. Les sons, dira-t-on, ont leurs propres intuitions et le poète est un « peintre aveugle » (Michel-Ange)."

back to all the words already uttered in the same portion of the speech (typically a verse or a sentence), but it cannot be compared to what it is still to come in the speech. In other words, the ear can go leftwards, not rightwards as the eye.

- As for detection of the alliteration I make no attempt to distinguish between **iconic** and **non-iconic** alliteration, which relates to style and may involve a certain level of arbitrariness (and for the same reason is also difficult to be understood by machine analysis).[12]

Without delving into the details, the algorithm I devised from the theoretical framework sketched out above acts on text portions defined on linguistical and metrical grounds, the sentence (in non-metrical texts) and the line (in metrical texts), which thus define its context of application. For instance, we will analyse a single line at a time, like the hexameter quoted above, *tympana tenta tonant palmis et cymbala circum.* According to the principles stated above, I start detecting alliterating sequences, defined as all the (not necessarily adjacent) "words" beginning with the same sound(s), which in this sample are *tympana tenta tonant* and *cymbala circum.* This approximation reflects the paramount role of initial sounds at the beginning of words as the building blocks of an alliterating sequence; it must be noted that not all the words need to be adjacent to each other, but typically at least a predefined number of them are (as we want to detect the repetition of initial sound(s) even when there is some other word between two words in the same context, but we do not want to treat as an alliterating sequence a sentence where e.g. the first and last word separated by several words happen to begin with the same sound): this number is a parameter for the scorer (a "sequence threshold"), typically set to 2; clients can customize it to change the scorer sensitivity to the accumulation of adjacent words. For each sequence I compare the initial sound(s) of each word to the initial sound(s) of each word to its left, picking up the pair with the longest portion in common and recording its length (the more sounds in common, the higher the effect) and relative distance (the nearer the words, the higher the effect).

This procedure takes into account the initial sounds of each word in the sequence as the basis for the repetition, according to the relevance of word-beginning sounds defined in the above theory and to the assumption that most of the ancient poetry is defined in acoustic rather than in visual terms, as it is orally performed for an audience. Thus the repetition of sounds becomes apparent once the same sound pattern is repeated by newly uttered words: the word just uttered can be traced back to all the words already uttered in the same portion of the speech (typically a verse or a sentence), but it cannot be compared to what it is still to come in the speech. I then apply at the level of the syllable in the context of a word the same procedure already applied at the level of the word in the context of a sentence (or verse), walking down the linguistic hierarchy: each syllable is compared with all the left syllables in the same word, but also with the

[12] For some examples see Herescu 108ff; for the discussion about *onomatopoeia* see also p.125 n.2.

first syllable of the first word of the sequence. This produces an array of counts of equal initial segments for each syllable, together with their relative distance, and constitutes what we can call the "syllabic heads" of a sequence. Finally, the segments of each syllables not included in the previous analysis (the syllabic heads) are also examined to see whether any of them are equal to the "master" segment in each sequence, which is simply the first segment of the first word in the sequence itself. For instance, in the sequence *tympana tenta tonant* the word *to.nant* has the second syllable containing the same /t/ segment beginning the sequence. The occurrence of this sound, even if not initial (which is already taken into account by syllabic heads), has somewhat the effect of "echoing" the dominant sound in the sequence and thus enforces its perception. This constitutes what we can call the "segmental echoes" of a sequence.

At this stage I have considered all the linguistic levels from "word" to syllable and phoneme, taking into account their position and distance, with a set of procedures which reflect the theoretical framework outlined above; once this analysis has been completed, I take into account the word heads, syllabic heads, and segmental echoes of every sequence in an input text (either a sentence or a line), combining them into a single numeric score using a set of parameters which assign a specific weight to several aspects of the algorithm, like word and syllabic heads distance and phonemic echoes; the greater the score, the higher the alliterating effect. Clients can change these parameters to variously adjust the scorer sensitivity to the texts being examined and the purposes of the analysis: typically scorer results on sample texts can be used at this stage for fine-tuning these parameters and then repeat the analysis several times until the best result is achieved.

To sum up, this sample starts from what might be considered a simple problem (detecting some repetition of sounds) to show the consequences of a trivial analysis ("counting letters"), which simply tries to take in all the available data with no preliminary evaluation of their relevance on theoretical grounds: a casual analyst might well be tempted by the raw computing power at his disposal to just count all the letters in every sentence of his texts to detect some alliterations in them. Of course, a more sophisticated analysis would refer to sounds rather than to letters, for the trivial fact that for historical and practical reasons no written language faithfully represents the corresponding sounds with a simple one–to–one ratio between letters and phonemes. In this case, a possibly complex preprocessing would be required to output e.g. a sequence of IPA characters from a sequence of a national alphabet letters. Even then (counting IPA characters rather than generic letters) the appreciation of the alliteration would probably be too inaccurate, as the analyst would just be facing a number of raw counts among which it would be very difficult to grasp what he can easily perceive by his ears: this is the situation already depicted by the pie-chart above, where each sound is given the same weight, while it should be clear that its position is the first factor defining the phenomenon under analysis. With such treatment probably the most emphasized alliterations might be detected, but we would also get a noise level so high that we might often be mislead in not only underestimating but (probably worse) overestimating several cases (remember Evan's samples like *Pergama Grais*).

In other words, not all the sounds have the equal weight for defining an alliteration, so several of them just represent noise data in a text: excluding them or at least reducing their weight it is not some sort of "trick" or adjustment we should take care of before feeding data to a machine, just because otherwise it would not "work" as expected; excluding them is right a consequence of a well-defined theoretical background, according to which it would be an error to do otherwise, an error which would *a priori* invalidate any experimental result, whatever it might be. We have seen that alliteration is a complex phenomenon and that in order to give at least a working definition of it, we must recur to linguistics and metrics, taking into account sentences, words, syllables, phonemes (rather than allophones), their "equality", positions and distances; and we cannot expect a machine to "understand" all these aspects without instruction, just feeding it with raw characters. It's up to our theory to define the relevance of each of these characters in the big picture we're trying to sketch, assigning them a well-rounded weight. Of course, it will probably be difficult to do this from the very beginning of our analysis: as a matter of fact, the results themselves will help us in fine-tuning all the parameters involved in our detection algorithm so that it "works" best for our texts and our purposes: we might want to give more, or less, relevance to the number of words which should be adjacent to define an alliterating sequence, to the count of phonemes that words or syllables share, to their relative distance etc., so that our analysis is sensitive to certain types of alliterations.

The same applies to the usage of neural recognition systems in contexts like palaeography: preprocessing here is not just a remedy for an otherwise poorly performing system. As we cannot expect miracles in detecting alliterations if we just feed a system with a bunch of letters, we can no more claim that a neural recognition system might be able to perform well with a relatively small sample where each shape is fed as it is, without instructing the system about the features scholars define as relevant for their data and purposes. As for the pie chart sampled above, we might well get some positive results even in the worst context, but the noise level and consequently the number of failures would be so high that such a system would be as unpractical for our purposes as the above chart is for detecting alliterations.

It must also be emphasized that the directions given to the system via preprocessing would probably be different according to the data being analysed and to our purposes in analysing them. The alliteration sample shows that different languages and literary traditions might well require a different adjustment of the various parameters, or even more significant changes in the algorithm itself: if we just refer to the theoretical points outlined above we can see that several aspects directly derive from various linguistical, metrical and literary assumptions (e.g. the role of suffixation in Indo-European morphology and its consequences for omeoteleuton and rhyme; the role of consonants vs. vowels in alliteration; the role of words, syllables, and phonemes in alliterating sequences; the relevance of word-initial sounds in poetic traditions like old German and ancient Latin; the importance of recitation in ancient poetry; etc.), which might have different weight in different languages, texts, literary genres, etc. Also, such software systems are created with specific purposes, and often it is important to adjust their sensitivity according to these purposes: for instance, running an alliteration detection system on a literary Italian corpus of prose texts where for some reason we might want to look for alliterating sentences would require a higher sensitivity than running it on a corpus of early Latin poetry.

The same applies to palaeographical applications, where our factors would rather be the types of shapes involved and their level of complexity and similarity, but also our purposes in setting up a recognition system: a system designed to recognize just different letters might probably require to be less sensitive than one designed to detect relevant differences between shapes of the same letter. If our data are not enough to provide a good number of samples we might even want to define some automatic distortion procedures which artificially increase them by applying to each letter shape some slight distortions which do not affect their relevant features. Even then, we would always refer to our theoretical framework, which would be the only judge for defining which features can be treated as relevant (and thus cannot be touched by the distortion process) and which can instead be freely altered, just like for alliteration our principles define which letters are relevant, and to what extent, and which represent just noise.

5 Digital Editions: Possible Approaches

On these premises, it seems very likely that such neural systems might find different types of application and success in different scenarios, and at any rate it is not possible to predict their outcome until they are tested “on the spot”, fine-tuning their parameters and eventually even the theoretical framework behind data preprocessing until we get the best results, according to the data type and the system purposes. As for any other technology, there is no miracle here, and such systems are orders of magnitude far behind the capacity of judgement of human beings: after all, we are not supposed to, nor would we probably want to, find some way of avoiding to think and set up a theoretical framework by ourselves. Such tools are just a convenient way of dealing with very large amount of data, and it is up to scholars to use them with judgement and profit. To sum up, the problem with palaeographical scenarios is that usually the most suggestive applications would require far too many samples to be practical or even possible, given our relatively limited set of data and/or the effort spent in their preprocessing: if we want our machine to recognize different letters like an OCR software we would have to lower our “similarity” threshold, so that the variants can be easily recognized as such; but if at the same time we also want to recognize chronological, regional, or personal variants of each different shape we would have to raise the threshold, maybe up to a level where the machine would be so picky that each shape would be judged as a different letter; finding the right balance would be very difficult if not impossible.

Rather, in a practical implementation we might think of partitioning the problems and limit the application of such techniques to pattern subsets, referring to a carefully defined theoretical background, just like in the alliteration sample. For instance, we might think of applications where visuals are even more compelling, like cataloguing brick stamps. In this case we’re dealing with letters of abbreviations which often are so connected and reshaped into monograms-like shapes that they can be treated as single drawings; of course, even with stamps it might often be the case that differences are too fine to be efficiently detected, but given their high number it could be useful to have a system trained only to recognize some single traits of each of the main classes, maybe just to provide users of a digital catalogue a quick and approximative way of detecting the major classification of a sample. In this context, an algorithmic approach describing each single shape would be extremely impractical, and even providing some sort of input for a hypothetical query system would be difficult. For example, once we have defined a subset of relevant shape traits, extracting them from the much more complex shapes of the actual samples we might think of presenting a relatively long list of them to the user to pick from; but this would probably be a cumbersome way of querying our catalogue, as users would have to carefully browse our samples list until they find what they are looking for. Even if the count of such more generalized samples were not so big, users would require a certain effort only to make a simple query.

A more user-friendly approach might be to provide users with a virtual board where they sketch with their mouse (or other pointing device) the traits we have defined as relevant for our super-classes, rather than looking for them in a relatively long list of prebuilt samples: then a neural network trained to recognize just these traits might be used to select the class of the sample being queried. This is a purely hypothetical example, but it shows how an ANN might even be used to just improve the users’ experience in defining their input for querying a digital catalogue: here the classes to be recognized would be much less than the actual stamps patterns, because editors would have restricted them to a much more abstract subset, defining only their most relevant traits. But this might be enough to restrict the query to a more significant set of classes (which might be later again restricted by means of similar or different query parameters, probably a combination of visual and non-visual ones), and this would be accomplished with a user simply drawing something, and letting the ANN recognize the most similar pattern in its trained set.

Here again it would be up to the scholars’ publishing the catalogue to define which traits should be selected as the most relevant partitioning criteria for their data, according to the theoretical framework they have set for their research method. Going one step further, this "partitioned" approach might also prove useful in palaeographical scenarios, where at a first approach it is not so easy to think of "decomposing" letter shapes into isolated traits which might not even make any sense for a human reader. Here a broader analogy might help: think of a typical identikit procedure, where witnesses must help building up the face of a person by selecting each trait from a set of galleries: they select head shape, nose, lips, ears, hairs, etc., one at a time from separated galleries of samples, picking one from each of them and then assembling all the components into a hypothetical portrait. As for letters, we have seen that in several cases we might have to face issues originated by the limited set of available samples in contrast with a high number of classes to be recognized, often distinguished by nuances which are too fine to be efficiently manipulated. We might think of partitioning the problem in advance, even before letting an ANN enter the scene, and define a set of single features we consider as relevant for distinguishing at least the most important classes. For instance, we might extract from letters just the shape of their serifs, the orientation of some selected segments, the shape of elliptical parts, etc., and define each as a separate step, to be treated by a specifically trained ANN. Like when using an identikit, users might just literally draw a sample for each of these features and let several trained ANNs recognize their patterns. Each of these steps would be a progress towards the final result, which would lead to some sort of classification of the letter sample. This way, the letter would no more be treated as a single pattern but rather deconstructed into a set of individual patterns, each recognized by a differently trained ANN. Later, we might be able to sum up the results of the output of all the ANNs, eventually combine them with other (non-visual) query parameters and get the desired classification. Of course, this is a hypothetical scenario, which would probably be practical only when dealing with a large amount of data and classes, or when no feasible alternatives to a purely visual way of querying a database of patterns are available; but it might be more rewarding than letting the machine try to perform a good recognition on data which are intrinsically very difficult to be treated.

Bibliography

Chang, Chih-Chung and Lin, Chih-Jen. *LIBSVM – A Library for Support Vector Machines.* <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.

Evans, Walter J. *Alliteratio Latina or Alliteration in Latin Verse Reduced to Rule with Special Reference to Catullus, Horace, Juvenal, Lucan, Lucretius, Martial, Ovid, Persius, Phaedrus, Priapeia, Propertius, Statius, Tibullus, and Virgil.* London: Williams and Norgate, 1921.

Fusi, Daniele. "Edizione epigrafica digitale di testi greci e latini: dal testo marcato alla banca dati." *Digital Philology and Medieval Texts*. Ed. Arianna Ciula, Francesco Stella. Pisa: Pacini, 2007. 121–163. <http://www.fusisoft.it/Doc/ActaArezzo.pdf>.

Fusi, Daniele. "An Expert System for the Classical Languages: Metrical Analysis Components." (to be printed in the Proceedings of the international conference *Trends in Computational and Formal Philology - An Italian Overview*. Venice and Padua, 22–24 May 2008). <http://www.fusisoft.it/Doc/ActaVenezia.pdf>.

Gershenson, Carlos. *Artificial Neural Networks for Beginners*. <http://arxiv.org/abs/cs.NE/0308031>.

Herescu, Nicolas I. *La poésie latine. Étude des structure phoniques*. Paris: Soc. d'Éd. Les Belles Lettres, 1960.

Monteil, Pierre. *La phrase relative en grec ancien*. Paris: Klincksieck, 1963.

Stergiou, Christos and Siganos, Dimitrios. *Neural Networks*. <http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html>.

Valesio, Paolo. *Strutture dell'allitterazione. Grammatica, retorica e folklore verbale*. Bologna: Zanichelli, 1967.

Approche informatique du document manuscrit*

Gilbert Tomasi, Roland Tomasi

Résumé

Les technologies mises en œuvre avec le logiciel BIT-Alpha, sont la base d'un outil informatique d'aide à l'analyse de l'écriture manuscrite naturelle, pour transcription et identification, pour une utilisation en paléographie. Dans l'image numérique, la capture du contenu du document en permet une analyse, puis une interprétation et une valorisation. La binarisation, la capture des lignes et la segmentation de l'image sont exposées et commentées. L'identification des mots, puis des lettres permet une première reconnaissance de l'écriture, basée sur l'analyse du graphisme. La transcription du texte s'appuie en complément sur des considérations linguistiques. Les critères d'analyse du graphisme permettent aussi une aide à l'identification du scribe. Une idée de mesure et de normalisation de la différentiation entre écritures est envisagée. Les éléments graphiques peuvent être édités. En exemple, le traitement d'un texte manuscrit est détaillé.

Zusammenfassung

Die in der Software BIT-Alpha eingesetzten Technologien sind die Grundlage der computergestützten Analyse, Identifizierung und Transkription von Handschriften sowie deren Interpretation und Bewertung in der paläographischen Forschung. Der vorliegende Aufsatz stellt die Verfahren der Binarisierung, der Zeilenerkennung und der Bildsegmentierung vor. Bei der graphischen Analyse des Schriftbildes, die eine erste Handschriftenerkennung erlaubt, werden zunächst die Wörter identifiziert und anschließend deren einzelnen Zeichen. Die Transkription des Textes stützt sich zusätzlich auf linguistische Methoden. Die zur Analyse des Schriftbildes herangezogenen Kriterien geben auch eine Hilfe zur Identifizierung des Schreibers. Die Erkennung eines mittelalterlichen handschriftlichen Textes wird im Detail an einem Beispiel dargestellt. Fernziel ist es, Messungen von Schriftzügen und eine normenbasierte Unterscheidung von Schrifttypen zu ermöglichen.

Abstract

The technologies used by the software BIT-Alpha are the basis as well for computer-aided analysis, identification, and transcription of writings as for their interpretation

[*] Nous désirons exprimer nos remerciements à Monsieur Torsten Schaßan de la Herzog August Bibliothek de Wolfenbüttel pour son soutien et conseil pour la réalisation de cet exposé.

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt : Books on Demand, 2009. 197–218.

and evaluation for palaeographical research. The present article presents methods for binarisation, line detection and image segmentation. During the graphical analysis of script which is used as initial approach to script recognition words will be recognised first and after them single characters. The transcription of text then has to be assisted by linguistic methods. The criteria which are drawn on for the analysis of script will help to identify individual scribes. The recognition of a medieval written text is presented in detail. Ideas of how differences between writing styles can be measured and normalized will present future prospects.

1 Introduction

La transcription informatisée de l'image numérique d'un texte en un texte utilisable informatiquement est réalisée communément par des logiciels dénommés OCR pour Optical Characters Recognition. S'il existe aujourd'hui des solutions OCR pour certains documents imprimés, les documents manuscrits restent globalement informatiquement inaccessibles, illisibles.

Sur des documents imprimés récents, la transcription informatisée d'un texte donne, avec l'utilisation de logiciels OCR usuels, des résultats utilisables, en particulier pour une recherche contextuelle. La transcription de documents imprimés anciens, avant 1850, constitue une difficulté demandant l'utilisation d'outils informatiques spécialisés, comme le logiciel BIT-Alpha, en mesure de lire tout document imprimé, indépendamment de l'époque et la langue. BIT-Alpha peut apprendre la fonte de caractères et mêmes les mots directement du texte à traiter. Ainsi BIT-Alpha peut transcrire par exemple des incunables, dont la fonte fut la copie de l'écriture manuscrite des livres anciens. Les technologies informatiques mises en œuvre avec BIT-Alpha, peuvent être la base d'un outil informatique permettant l'analyse de l'écriture manuscrite et constituer une aide à la retranscription de l'écriture manuscrite naturelle. Un logiciel OCR spécialisé sur l'écriture manuscrite est aussi appelé ICR pour Intelligent Characters Recognition, car il est supposé reconnaître des caractères générés par l'intelligence et non par une machine.

C'est cette approche informatique du document manuscrit que nous voudrions exposer ici en présentant notre développement, le logiciel BIT-Alpha_ICR, comme outil d'analyse de l'écriture et d'aide à la retranscription, précisément dans la perspective d'une utilisation en paléographie. Dans l'image numérique, il s'agit en premier de capturer le contenu du document, pour en faire ensuite une analyse, puis une interprétation et une valorisation. La lecture informatisée d'un document suppose donc primairement la capture du contenu de ce document.

Au départ, il s'agit de rendre l'image d'archive, souvent très lourde car en couleur et de haute définition, utilisable informatiquement pour un traitement ICR. Cette première transcription, appelée binarisation, crée une image binaire et voudrait éliminer le bruit de fond du scanner, les taches et colorisation rendant toute lecture impossible, et cela si possible en gardant le détail de l'écriture, ponctuation et point sur les i ! C'est loin d'être une problématique simple et elle est souvent négligée, alors qu'il s'agit de la base de toute analyse et retranscription informatisée.

Ensuite une segmentation de l'image est nécessaire, pour distinguer dans l'image du document, des régions contenant un texte et celles contenant un élément graphique, afin de les découper pour les traiter séparément et différemment. La segmentation va de paire avec le redressement de l'image pour placer les lignes du texte le plus horizontalement possible.

Une des difficultés majeure du traitement de l'écriture manuscrite constitue en la capture des lignes du texte, celles-ci n'étant naturellement pas droites ou alors soulignées d'un trait de ligne, gênant la reconnaissance et de plus souvent les lignes sont imbriquées, les lettres se chevauchant et se touchant. Une fois le texte isolé et les lignes déterminées, suit l'analyse du contenu en vue d'une transcription. Il s'agit pour chaque ligne de texte, d'en capturer les mots, puis de capturer les lettres et de proposer une reconnaissance de l'écriture.

Nous procéderons en reconnaissance ICR par une double approche, partant en premier de la capture et de l'analyse globale du mot, capturé en tant qu'unité graphique à analyser dans son tout, en recherchant une corrélation avec un mot d'un vocable connu dont le graphisme a été appris au préalable, un peu comme la méthode de lecture dite « globale ».

Ensuite, cette identification globale du mot sera confrontée à une analyse des lettres constitutives du même mot. La capture individuelle des lettres suppose la séparation des caractères entre eux, ce qui constitue une vraie difficulté dans le cas de l'écriture liée naturelle. Avec la capture et l'analyse des caractères, pourra être envisagée leur identification en vue de proposer une transcription. Celle-ci est basée sur divers critères corrélés à une analyse de forme. Une fois les éléments constitutifs du texte analysés, il sera procédé à leur interprétation pour valorisation. Pour être pertinente, la transcription du texte devra s'appuyer sur des considérations linguistiques en complément aux informations de source purement graphique.

Les critères d'analyse des lignes du texte, des mots et des caractères permettront aussi une aide à l'identification du scribe. Une idée de mesure et de normalisation de la différentiation d'écritures est envisagée. Enfin les éléments graphiques peuvent être édités et constituer une base de données permettant aussi d'identifier le scribe ou servir une étude spécifique.

En résumé BIT-Alpha_ICR constitue un outil permettant l'analyse de l'écriture manuscrite et sa retranscription avec l'aide d'un computeur. Une quantification normée de la différentiation d'écritures serait d'un grand intérêt pour la paléographie. En perspective, une application vers un traitement ICR automatisé peut être envisagée ainsi que la lecture ICR de textes en écriture liée comme l'arabe ou l'hindi.

2 Capture du contenu

Préalablement, il convient de lever un éventuel malentendu : nous voulons traiter informatiquement l'écriture manuscrite de documents anciens ou dits « existants », c'est-à-dire partant de l'image numérique, généralement scannée, d'un document existant sous forme papier ou équivalent. Il s'agit donc d'un traitement d'image et de l'analyse du tracé d'une écriture « existante ». Cette démarche ne doit pas être confondue avec la reconnaissance de l'écriture manuscrite « vivante » obtenue par capture du mouvement au moment même de l'acte d'écrire et de la génération de l'écriture, par exemple sur un capteur graphique, comme une tablette graphique ou un écran tactile. Dans ce cas, sont aussi disponibles pour l'analyse informatique des informations dynamiques, comme la vitesse du tracé, l'ordre d'exécution des différents traits, la force d'appuis, etc. Autant d'éléments dynamiques permettant une reconnaissance finalement plus aisée et communément utilisée aujourd'hui sur certains appareils portables. Dans ce cas d'analyse d'informations dynamiques, il est usuel de parler de reconnaissance « on-line » en opposition à l'analyse d'image d'un texte scanné, appelée reconnaissance « off-line » et qui constitue ici notre propos.

En partant d'une image numérique d'archive, souvent de haute définition, typiquement de 300 à 600 dpi couleurs, il s'agit en premier d'obtenir une image utilisable pour un traitement de reconnaissance de caractères. Au final, il faut obtenir une image binaire et d'une définition de 300 à 400 dpi, une définition moindre rendant la distinction impossible entre des caractères proches et une image de trop haute définition faisant de chaque caractère une pièce unique.

2.1 La binarisation

Le procédé de binarisation constitue une opération informatique de préparation au traitement de reconnaissance et qui n'est pas toujours intégrée dans un logiciel OCR ou ICR. Elle constitue pourtant une étape majeure dans l'analyse et demande grand soin, car une mauvaise binarisation peut détruire beaucoup d'informations contenues dans le facsimilé scanné de l'original. Par contre la binarisation doit rendre le document utilisable pour une reconnaissance et doit pour cela effectuer des opérations délicates : supprimer le bruit de fond du scanner, tenir compte des variations de contraste et de luminosité sur la page scannée, éliminer les taches gênantes, nettoyer les bords et zone de marges sans texte, obtenir le plus de détails possibles dans la zone texte, c'est-à-dire éliminer la face arrière du document visible par transparence ou parce que l'encre à diffuser dans le papier, mais garder la ponctuation et les points sur les i.

A côté d'autres outils de binarisation plus usuels, comme un seuillage de luminosité ou de couleur, mis aussi en œuvre dans BIT-Alpha mais avec la particularité d'une localisation par région de texte, c'est-à-dire différent en haut et en bas de la même page, le

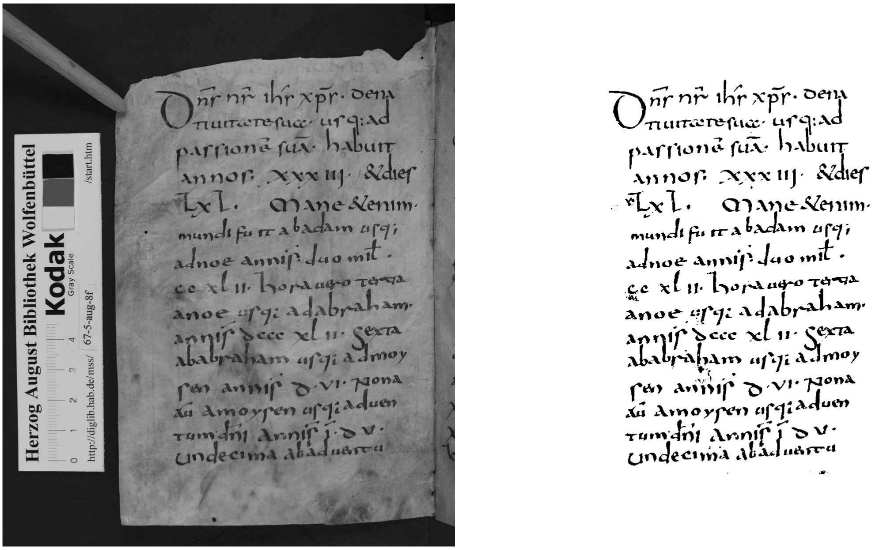

Herzog August Bibliothek Wolfenbüttel
Kodak
Gray Scale
http://diglib.hab.de/mss/ 67-5-aug-8f /start.htm

FIG. 1. Un document numérique et sa binarisation par BIT-Alpha. Wolfenbüttel, Herzog August Bibliothek, Cod. Guelf. 67.5 Aug. 8°, fol. 2v

logiciel BIT-Alpha propose aussi une binarisation utilisant l'algorithme de Niblack, modifié par nos soins pour en corriger certaines imperfections, afin de satisfaire au mieux les impératifs contradictoires de la binarisation. Cet algorithme reconnu comme un des mieux adaptés pour ce faire, procède d'une analyse dit « dynamique » de contraste local et cela individuellement autour de chaque caractère. De plus une analyse de la répartition gaussienne des contrastes, permet de distinguer les zones textes, des zones de marge, afin de les traiter avec des seuillages différents. Ce procédé génère une image binaire de bonne qualité et c'est le type de binarisation que nous préconiserons pour le traitement de documents manuscrits. On y ajoutera éventuellement un prétraitement d'image permettant d'éliminer une couleur correspondant à une tache gênante, comme la coloration intempestive en rouge de certaines majuscules, ce qui les rendrait illisibles informatiquement.

La figure 1 montre un exemple de binarisation avec l'algorithme de Niblack modifié, tel que mis en œuvre dans BIT-Alpha.

2.2 La segmentation

Après avoir nettoyé et réduit l'image d'archive en une image binaire plus légère et plus facilement traitable informatiquement, il est nécessaire de distinguer dans cette image les zones contenant un texte de celle contenant une image ou des éléments graphiques. Cette séparation de l'image en différentes régions est appelée communément segmentation.

Dans le cas d'un outil d'aide à la transcription d'un texte manuscrit où l'intervention humaine est prévue pour finaliser la transcription, cette étape peut paraître secondaire dans la mesure où l'opérateur peut lui-même isoler la zone texte l'intéressant. Malgré tout, elle est d'intérêt en permettant d'isoler et par là de capturer, les différents éléments graphiques. Dans le futur, un procédé entièrement automatique ne pourrait se concevoir sans segmentation, telle qu'elle est mise en œuvre aujourd'hui pour des livres anciens ou des journaux par BIT-Alpha.

La segmentation va souvent de paire avec l'orientation de la page, rotation dénommée aussi redressement, afin d'obtenir du mieux possible un texte aux lignes plus ou moins horizontales, ce qui simplifie grandement la capture des lignes et la reconnaissance des caractères. Concernant un texte manuscrit, le redressement sera de toute autre difficulté si nous avons affaire à un texte écrit bien droit, avec éventuellement des lignes guide que suit l'écriture manuscrite, comme c'est souvent le cas, ou un texte aux lignes courbes et écrit « en vrac ».

Pour le redressement d'un texte manuscrit, nous préconiserons un procédé suivant les travaux de Frank Le Bourgeois de l'I.N.S.A. Lyon et partant d'une analyse différentielle de la variance bidirectionnelle dans l'image, telle que décrite dans la publication citée en bibliographie.

2.3 La capture des lignes

La capture des lignes d'un texte manuscrit est la première difficulté majeure. En effet, même si les lignes de texte suivent un trait du document, les caractères sont souvent très imbriqués et se touchent d'une ligne à l'autre. Parfois un document manuscrit présente des lignes courbes, ce qui complique le traitement informatique. Un procédé usuel donnant de bons résultats sur des textes imprimés aux lignes clairement séparées et rectilignes, ne conviendra pas aux variations de l'écriture manuscrite. BIT-Alpha met en œuvre pour la capture des lignes les travaux de recherche du Laboratoire CNRS L.I.R.I.S. de I.N.S.A. Lyon autour du Professeur Hubert Emptoz (Le Bourgeois et al. 2004) et concernant aussi des applications en paléographie (Emptoz).

Pour la capture des lignes du texte, il est procédé à une analyse différentielle de la variation unidimensionnelle dans l'image et permettant d'obtenir les zones constitutives des lignes du texte, comme visible dans l'image figure 3. Cette méthode mathématique

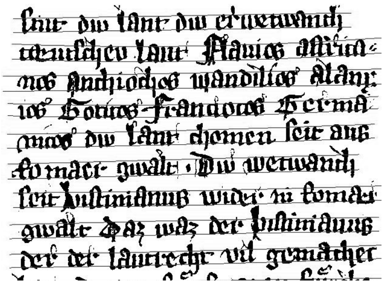


FIG. 2. Capture de lignes dans un texte manuscrit. Wolfenbüttel, Herzog August Bibliothek, Cod. Guelf. 15.2 Aug. 2°, fol. 1v

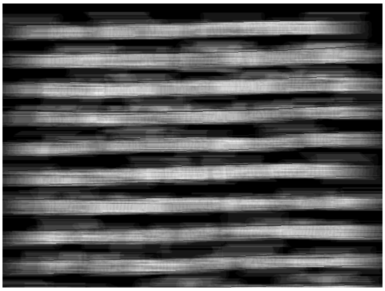


FIG. 3. Analyse différentielle de l'image du texte manuscrit de la figure 2.

est la base du procédé de reconnaissance de lignes de BIT-Alpha, laquelle a été complétée et adaptée pour tenir compte de différentes difficultés inhérentes à la méthode, comme la capture des grandes lettres d'un titre ou de symboles isolés. Cette méthode a le gros avantage de bien reproduire la courbure des lignes de texte et fonctionne indépendamment de l'identification de caractères. Après avoir déterminé la zone constitutive d'une ligne, encore faut il en déduire, à partir d'une suite de points, une fonction mathématique donnant la ligne elle-même. Ce procédé que nous jugeons très performant constitue la quatrième génération de procédés de capture de lignes mis en œuvre dans BIT-Alpha et se montre supérieur aux autres méthodes précédemment utilisées dans BIT-Alpha et partant soit des caractères eux-mêmes, soit d'une analyse de périodicité ou de projection d'ombres. Ces derniers procédés ne pouvant convenir à l'écriture manuscrite qui n'a en général ni caractères séparés et homogènes ni périodicité établie.

Ce procédé d'analyse différentielle a l'avantage, non seulement de permettre de capturer les lignes, mais aussi d'obtenir la base-ligne, et la top-ligne comme visible dans l'image figure 2. Cette information, de première importance pour la reconnaissance individuelle des caractères, donne aussi accès, avec la distance moyenne entre ces lignes, à une grandeur de référence pour la taille de l'écriture, premier élément pouvant caractériser une écriture manuscrite déterminée.

3 L'analyse du contenu

Une fois le texte isolé et les lignes du texte déterminées, il faut procéder à l'analyse des mots et des caractères pour obtenir une identification permettant l'accès au texte, au contenu. Cette identification servira de base à une première proposition de transcription du texte, basée uniquement sur l'analyse de l'image.

3.1 Analyse des mots

Une fois les lignes du texte déterminées, nous capturerons les mots de chaque ligne pour une première analyse d'identification. Alors que les OCR procèdent généralement pour reconstituer les mots d'une approche partant de la capture individuelle des lettres constitutives, séparées entre elles et réputées identiques dans les imprimés, un ICR est confronté à des lettres souvent liées entre elles et présentant une forte variance graphique. De ce fait, une approche globale considérant en premier le mot entier dans son ensemble s'impose au prime abord.

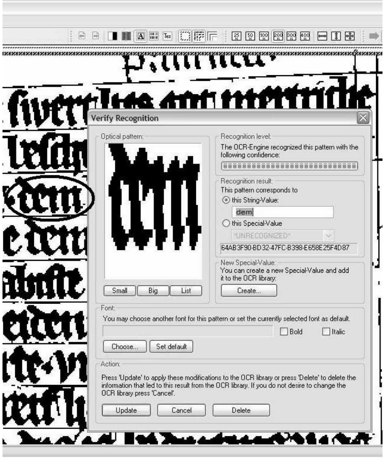
La capture des mots sera facilitée dans un texte à l'écriture liée ou chaque mot constitue naturellement un ensemble graphique séparé les uns des autres. La chose sera plus délicate dans un texte ou les lettres manuscrites ne seront pas toutes liées et avec une distance entre elles changeante. Une analyse de la distance entre les caractères devrait permettre une première capture des mots. Ici BIT-Alpha dispose d'outil d'analyse assez fin mis au point pour lire des textes imprimés de la Renaissance ou des journaux dans lesquels la distance entre lettres et mots est souvent très fantaisiste et n'est pas plus homogène que dans un texte manuscrit. Cette analyse de la distance entre lettres a lieu individuellement pour chaque ligne.

La figure 4 montre un exemple de capture et d'identification après l'apprentissage d'un mot entier dans BIT-Alpha. Il s'agit du mot « dem » qui est en premier appris en tant qu'un tout. Dans la figure 5, un autre mot « dem » est ainsi reconnu dans une autre partie du texte, affichant un taux de reconaissance moindre.

C'est une particularité de BIT-Alpha de pouvoir, dans une première phase d'apprentissage, enregistrer l'identification donnée par l'opérateur et apprendre, directement du document à traiter, le graphisme des éléments à reconnaître. BIT-Alpha est un OCR du type « self-learning » ou « adaptatif » procédant par une phase d'apprentissage pour reconnaître ensuite les symboles appris au préalable. Ceci est valable pour les mots dans leur globalité graphique et pour les symboles individuellement.

3.2 Analyse des caractères

Pour obtenir une reconnaissance fiable, l'analyse des mots dans leur tout ne suffit pas, il faut compléter par une analyse des caractères constitutifs du mot afin de corréler ces deux informations. Cette analyse suppose de pouvoir capturer les caractères individuellement, c'est-à-dire de pouvoir les considérer isolément. Ceci suppose qu'ils



Verify Recognition

Optical pattern:

Recognition level:
The OCR-Engine recognized this pattern with the following confidence:

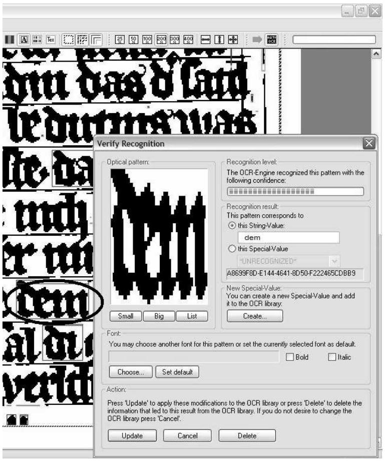
Recognition result:
This pattern corresponds to
● this String-Value: dem
○ this Special-Value *UNRECOGNIZED*
64AB3F90-BD32-47FC-B398-E658E25F4D87

New Special-Value:
You can create a new Special-Value and add it to the OCR library:

Font:
You may choose another font for this pattern or set the currently selected font as default.
□ Bold □ Italic

Action:
Press 'Update' to apply these modifications to the OCR library or press 'Delete' to delete the information that led to this result from the OCR library. If you do not desire to change the OCR library press 'Cancel'.

FIG. 4. Capture et apprentissage du mot « dem ».



Verify Recognition

Optical pattern:

Recognition level:
The OCR-Engine recognized this pattern with the following confidence:

Recognition result:
This pattern corresponds to
● this String-Value: dem
○ this Special-Value *UNRECOGNIZED*
A8699F8D-E144-4641-8D50-F222465CDBB9

New Special-Value:
You can create a new Special-Value and add it to the OCR library:

Font:
You may choose another font for this pattern or set the currently selected font as default.
□ Bold □ Italic

Action:
Press 'Update' to apply these modifications to the OCR library or press 'Delete' to delete the information that led to this result from the OCR library. If you do not desire to change the OCR library press 'Cancel'.

FIG. 5. Reconnaissance d'un autre mot « dem » dans une autre partie du texte.

soient séparés les uns des autres ou que nous disposions d'un procédé pour séparer les caractères collés. Ce procédé que nous avons nommé « disjonction » est intégré dans BIT-Alpha.

Ici aussi, après plusieurs premières générations d'algorithme ou la disjonction avait lieu suivant des critères géométriques, le procédé de disjonction actuellement utilisé dans BIT-Alpha part d'une approche « d'intelligence artificielle » en séparant les caractères lorsque le logiciel ICR reconnaît les composants à séparer. Cette démarche, plus complexe à mettre en œuvre que son simple énoncé, donne dans la pratique d'excellents résultats, confirmés en production de masse avec BIT-Alpha. Nous sommes de l'avis que ce soit le procédé unique ouvrant l'accès à la reconnaissance ICR devant traiter l'écriture naturelle liée.

Dans la reconnaissance ICR d'un fragment de texte, comme un code postal et sa ville ou un nombre en toutes lettres, il est d'usage de procédé d'une démarche atomisant le mot en une suite de traits verticaux pour reconstituer ensuite les « m », « u », « n » et « a » en quelque sorte en comptant le nombre de pieds. Cette approche que nous avons aussi testée, nous semble inférieure à une reconnaissance du caractère dans son intégrité, basée sur la disjonction après reconnaissance ICR du caractère à isoler.

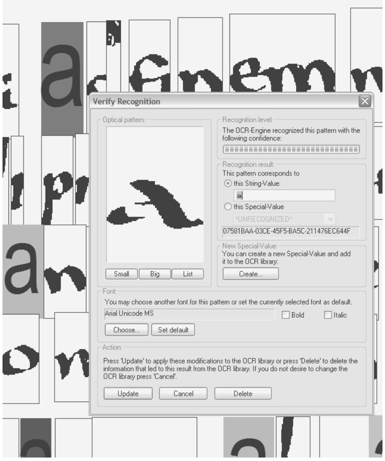
Notons au passage que si un ICR fonctionne sur un fragment de texte assez court et dont les éléments à reconnaître sont « attendus », c'est-à-dire connus et présents dans une liste de villes en référence ou des chiffres déjà lus, la difficulté est toute autre lorsqu'on se trouve confronté à la lecture d'une page d'écriture manuscrite dont le contenu est inconnu. Concernant les documents manuscrits historiques, la problématique est donc sensiblement différente, dans la mesure où il s'agit de reconnaître plusieurs pages d'un texte au contenu inconnu, mais écrit par une seule et même personne, appelée ici scribe, et dont l'écriture est apprise préalablement et non, à la différence, d'un texte limité et dont le contenu est connu ou du moins attendu, mais par contre écrit par des scribes inconnus et variés, comme c'est le cas pour la reconnaissance ICR par exemple d'adresses postales ou de nombres sur des chèques.

Profitons de cette parenthèse pour souligner que nous n'avons pas la prétention d'avoir une solution automatique pour la lecture ICR de page manuscrite inconnue, un domaine aux difficultés toujours non résolues, mais nous voulons simplement ici présenter notre approche et l'état de notre technologie. Le logiciel BIT-Alpha_ICR pouvant constituer un outil d'aide à la transcription ou d'aide à l'identification du scribe d'un texte. BIT-Alpha offre la possibilité, dans une première phase d'apprentissage, par un simple clic, d'apprendre un élément graphique isolé, tel qu'un caractère.

Les images figures 6 et 7 montrent la capture d'une lettre « a » dans BIT-Alpha, l'apprentissage de cette lettre et sa reconnaissance dans plusieurs mots et après disjonction dans trois autres mots du texte. Dans la figure 6 la lettre « a » est apprise, c'est-à-dire que l'opérateur en a donné la transcription, et elle est donc reconnue à 100 %, ce qui est indiqué par la couleur bleue (voir aussi figure 12 pour les couleurs). Par exemple, le « a » de « ra » qui n'est pas lié, est alors aussi reconnu, sans avoir été appris, avec une haute fiabilité, ce qu'indique la couleur verte. La disjonction n'est pas activée. Les « a » de « ad » et « am » qui sont liés ne sont pas reconnus. La figure 7 montre, après activation de la disjonction, la séparation de la lettre « a » de « ad » et « am » et la reconnaissance individuelle de la lettre « a » séparée du reste du tracé graphique.

Ainsi, il sera possible dans un texte manuscrit où les caractères sont collés entre eux, d'obtenir par disjonction une capture individuelle des caractères, et leur identification afin de reconstituer le mot. Cette méthode de disjonction suppose donc la reconnaissance des symboles à séparer, donc aussi un apprentissage préalable des caractères constitutifs des mots. Dans une écriture où tous les caractères seraient liés, la mise en œuvre de la disjonction suppose alors que lors de la phase d'apprentissage, l'opérateur puisse définir à quel endroit du graphisme du mot il convient de séparer les caractères afin de pouvoir apprendre le symbole individuellement. Cette fonction de disjonction manuelle de lettres collées, spécifique à un ICR, sera implémentée dans BIT-Alpha_ICR.

D'une façon plus générale, nous devons ici insister sur la spécificité du logiciel « self-learning » BIT-Alpha qui part en premier d'une démarche d'apprentissage des éléments à reconnaître : BIT-Alpha ne va reconnaître que le type d'écriture qu'il a auparavant appris. Il est évident que son taux de fiabilité de lecture est alors supérieur a celui d'un OCR « omni-fonte » supposé connaître toutes les types d'écriture et risquant de les confondre toutes. Les différentes bibliothèques de bases de reconnaissance OCR contenant les « empreintes digitales » des différents symboles et graphismes appris par



Verify Recognition

Optical pattern:

Recognition level:

The OCR-Engine recognized this pattern with the following confidence:

Recognition result:

This pattern corresponds to

● this String-Value: a

○ this Special-Value: *UNRECOGNIZED*

07581BAA-03CE-45F5-BA5C-211476EC644F

New Special-Value:

You can create a new Special-Value and add it to the OCR library:

Font:

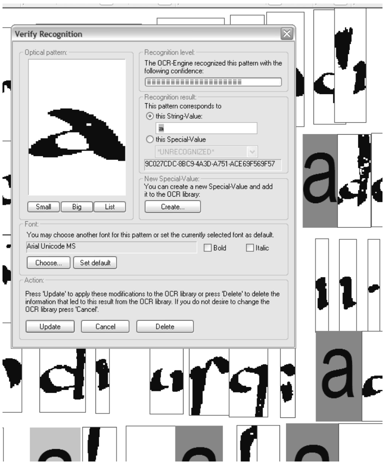
You may choose another font for this pattern or set the currently selected font as default.

Arial Unicode MS □ Bold □ Italic

Action:

Press 'Update' to apply these modifications to the OCR library or press 'Delete' to delete the information that led to this result from the OCR library. If you do not desire to change the OCR library press 'Cancel'.

FIG. 6. Lettre « a » apprise (au centre) avec la fenêtre de dialogue d'apprentissage ouverte, d'autres « a » non liées sont aussi reconnues.



Verify Recognition

Optical pattern:

Recognition level:

The OCR-Engine recognized this pattern with the following confidence:

Recognition result:

This pattern corresponds to

● this String-Value: a

○ this Special-Value: *UNRECOGNIZED*

9C027CDC-8BC9-4A3D-A751-ACE69F569F57

New Special-Value:

You can create a new Special-Value and add it to the OCR library:

Font:

You may choose another font for this pattern or set the currently selected font as default.

Arial Unicode MS □ Bold □ Italic

Action:

Press 'Update' to apply these modifications to the OCR library or press 'Delete' to delete the information that led to this result from the OCR library. If you do not desire to change the OCR library press 'Cancel'.

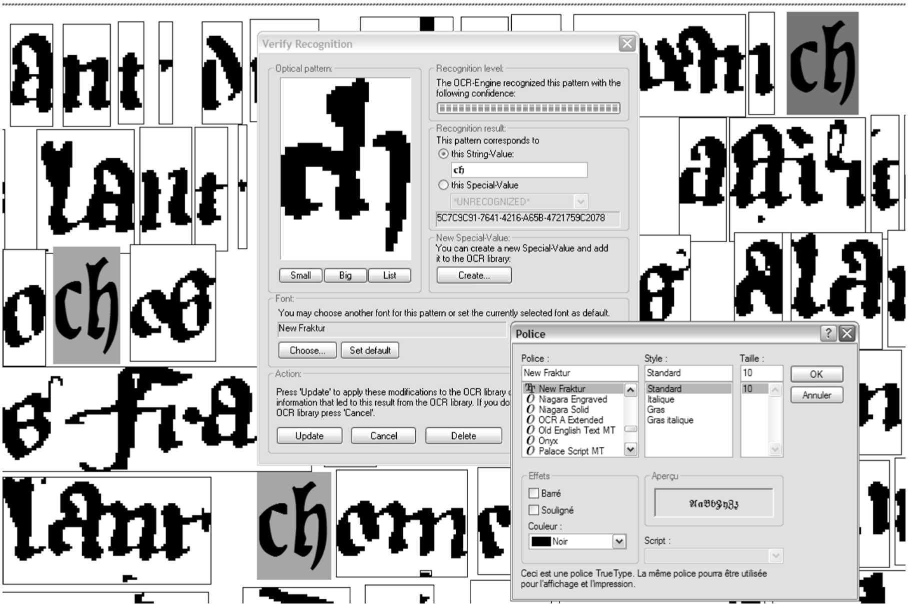
FIG. 7. Autres lettres « a » liées, reconnues une fois la disjonction activée.

BIT-Alpha, sont individuellement enregistrées. Cela permet de les inter-changer pour contrôle individuel ou de les mixer pour constituer une base plus générale. Il est ainsi possible d'identifier une écriture particulière en n'utilisant qu'une base OCR spécifique ne « connaissant » que l'écriture particulière d'un scribe connu afin de vérifier si le texte est bien de ce scribe. Ou bien, au contraire de rassembler en une seule base plus large toutes les écritures apprises afin de transcrire au mieux un texte inconnu.

BIT-Alpha peut apprendre et reconnaître aussi des ligatures comme le « ch » de l'exemple figure 8. Mais BIT-Alpha peut aussi apprendre et donc reconnaître toute suite des caractères convenant à l'analyse du texte.

BIT-Alpha dispose aussi de la possibilité, à l'inverse, de recoller des symboles disloqués, comme le « h » en bas de la figure 6. Exposer ici le détail de cette procédure de « séquenceur » serait fastidieux. Disons qu'elle part simplement du principe de séquences de symboles, par exemple qu'un « l » suivi d'une partie droite de « petit h » donne un « h ». Ces séquences sont individuellement programmables dans BIT-Alpha suivant les nécessités du texte et sont intégrées dans les bibliothèques ICR interchangeables.

Bien plus, la reconstitution des symboles à partir des éléments graphiques constitutifs pour établir une capture pertinente, est dans BIT-Alpha individuellement programmable. Il est ainsi possible avec BIT-Alpha d'obtenir par exemple dans un texte Fraktur



Verify Recognition

Optical pattern:

Recognition level: The OCR-Engine recognized this pattern with the following confidence:

Recognition result: This pattern corresponds to

● this String-Value: cħ

○ this Special-Value: *UNRECOGNIZED*

5C7C9C91-7641-4216-A65B-4721759C2078

New Special-Value: You can create a new Special-Value and add it to the OCR library:

Font: You may choose another font for this pattern or set the currently selected font as default.

New Fraktur

Action: Press 'Update' to apply these modifications to the OCR library information that led to this result from the OCR library. If you do OCR library press 'Cancel'.

Police

Police : New Fraktur

Style : Standard

Taille : 10

New Fraktur
Niagara Engraved
Niagara Solid
OCR A Extended
Old English Text MT
Onyx
Palace Script MT

Standard
Italique
Gras
Gras italique

Effets

□ Barré

□ Souligné

Couleur : Noir

Aperçu

Script :

Ceci est une police TrueType. La même police pourra être utilisée pour l'affichage et l'impression.

FIG. 8. Exemple d'apprentissage de ligature, ici « ch ».

imprimé ou pour des coursives de la Renaissance, la séparation de « e » ou « o » sous un « s long » tout en recollant des « n » ou « u » disloqués en petits traits verticaux. Les mêmes possibilités existent pour l'écriture manuscrite et le lecteur attentif remarquera que dans la figure 6 le « h », en bas à droite de page, et capturé disloqué alors que la même « h » est capturé intégralement dans la figure 11.

3.3 Analyse du texte

L'analyse d'image nous donne donc pour chaque mot deux identifications :

1. Suite à l'analyse globale des mots :
   La proposition d'un mot, en comparaison aux graphismes de mots entiers, auparavant appris en tant qu'entités géométriques et en mémoire dans une bibliothèque de bases ICR « d'empreintes numériques » de mots.
2. Suite à l'analyse des caractères constitutifs des mots :
   La proposition d'un mot reconstitué par la juxtaposition dans le bon ordre de ses caractères identifiés et reconnus par rapport à une bibliothèque de bases ICR « d'empreintes numériques » de caractères, aussi auparavant appris.

Dans la mesure où BIT-Alpha_ICR pourra de proposer le premier et le deuxième ou troisième choix, une corrélation de ces deux informations devrait permettre une première proposition de transcription du mot à lire informatiquement. A ce stade, tout ce que l'on a pu obtenir comme information pour la transcription du texte vient essentiellement de l'analyse du graphisme, du tracé du texte.

4 Valorisation du contenu

Pour valider la transcription, il faut maintenant faire intervenir des considérations basées sur le langage.

Une fois les lignes, les mots et les lettres capturés et analysés pour obtenir une proposition de mots pour la transcription du texte, celle-ci doit être affinée en corrélation au langage utilisé par l'auteur du texte. Cela suppose d'en connaître effectivement la langue et de disposer d'un lexique de mots ou d'un texte correspondant. Les abréviations pourront être transcrites avec le symbole Unicode correspondant au mieux au symbole original ou bien en les remplaçant par leur forme développée. Enfin BIT-Alpha propose plusieurs possibilités pour l'identification de la personne ayant écrit le texte, désignée ici par scribe.

4.1 Transcription du texte

En combinaison l'analyse du mot dans son ensemble et celle des lettres constitutives de ce mot, le logiciel propose un choix de mots pour la retranscription du texte.

Cette analyse du graphisme propose donc un mot comme résultat d'une analyse ICR purement basée sur le tracé de l'écriture. Ces considérations géométriques seront complétées par des aspects linguistiques de corrélation en référence à une liste de mots connus de la langue du texte à déchiffrer, afin de permettre de restreindre le choix et de valider la transcription proposée. Il sera aussi possible de recoller ainsi des lettres isolées en début ou en fin de mot.

Le premier résultat de lecture ICR va donc devoir être confronté à des considérations linguistiques pour le valider et permettre une identification et proposer une transcription du texte la plus correcte possible. La langue du texte analysée étant supposée connue, le mot proposé par l'ICR, ou les premier, second et troisième choix de mots proposés, vont être corrélés à un lexique de mots de la langue utilisée afin d'en valider le bon choix pour le mot transcrit. Dans BIT-Alpha_ICR nous mettons en œuvre l'évaluation de la distance de Levenshtein (Ringlstetter). Cette distance est une grandeur donnant la mesure du coût pour modifier un mot en un autre mot. BIT-Alpha_ICR est capable de gérer des bases de données considérables, comme par exemple un lexique de 500 000 mots latins.

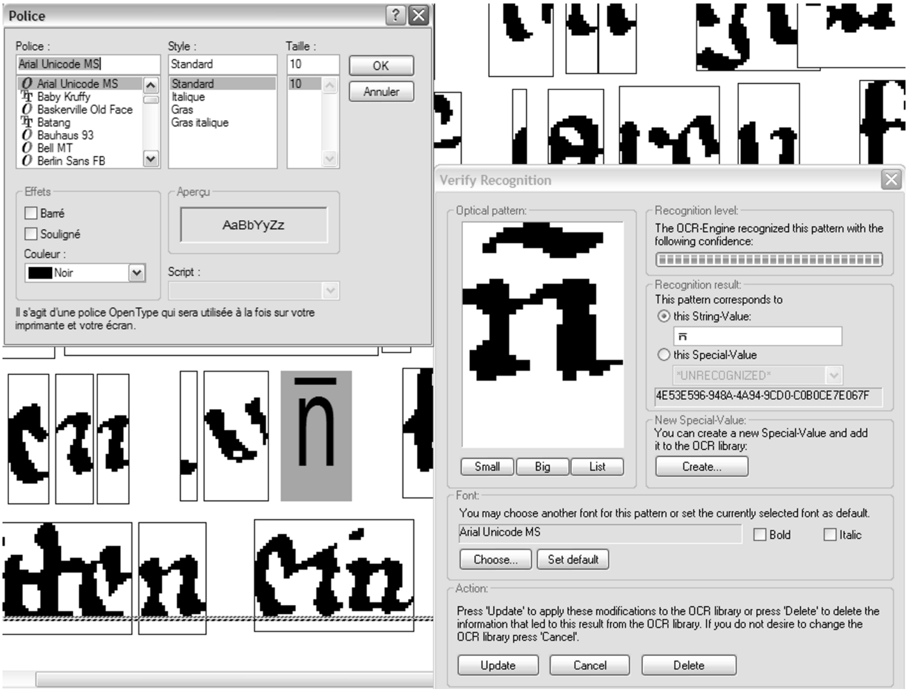
Mais nous sommes d'avis que ces seules considérations de correction de textes ne suffisent pas dans le cas de la validation d'un mot lu par ICR. En effet il ne s'agit pas de corriger des fautes de frappe par exemple, ni des fautes d'orthographe, mais bien plus, des fautes de lecture ICR. Il nous semble donc nécessaire d'ajouter aux considérations de corrélation basique, une pondération basée sur la probabilité d'erreur de lecture ICR, c'est-à-dire la probabilité de confusion entre deux lettres proches. Par exemple en reconnaissance de l'écriture Fraktur ou en Roman de la Renaissance, le «ſ» long sera souvent confondu avec un «f». Concrètement par exemple un mot d'un texte français lu «ſut» devra plutôt être corrigé « fut » que « sur ». Cette pondération de la correction linguistique nous paraît essentielle en lecture ICR et nous l'avons implémentée dans BIT-Alpha en considérant aussi les travaux de Frank Le Bourgeois et Hubert Emptoz du laboratoire CNRS I.R.I.S. de l'INSA Lyon (Emptoz et al.). Ceci permet à l'opérateur de définir pour une bibliothèque de mots donnée, une matrice de coefficients de corrélation entre les lettres sujettes à confusion. Comme pour tout dans BIT-Alpha l'opérateur maîtrise tous les paramètres et décide lui-même de ce que doit effectuer ou pas le logiciel.

Notons ici pour mémoire que le « séquenceur » déjà expliqué au chapitre 2.3 permet d'implémenter des séquences typiques de la langue, donnons à titre d'exemple et simplement pour la compréhension du principe : « nnn » donne « mm ». C'est un instrument très puissant de correction d'erreurs de lecture et peut ainsi permettre une meilleure transcription du texte.

Dans la mesure où la langue du texte est connue et où un échantillon de texte (transcrit en « texte mode » utilisable informatiquement) de cette langue est disponible, une analyse de la codification du texte permettrait aussi de restreindre le choix des lettres convenant. Il s'agit ici d'une méthode utilisée communément en déchiffrage et prenant en considération la fréquence d'une lettre, ses appareillages avec d'autres lettres (syllabes usuelles) et aussi la statistique de répartition de sa position dans les mots, début, fin, milieu. Théoriquement ces seules considérations permettraient en partant de l'analyse de codification d'une seule page de texte de la même langue (en texte mode), d'identifier automatiquement les lettres d'un texte inconnu mais de cette même langue et donc d'en faire un apprentissage entièrement automatisé, sans intervention humaine.

Tous ces aspects peuvent être l'assise d'une correction, basée sur la langue, de la première lecture ICR basée sur l'analyse du graphisme, et aboutir aussi à une première proposition de texte dans le cadre d'une aide informatisée à la transcription.

Dans les textes manuscrits patrimoniaux, les auteurs utilisent très souvent des abréviations afin d'écrire plus vite. La retranscription du texte sous une forme lisible par un non spécialiste demande la transcription de ces abréviations. Pour pouvoir transcrire une abréviation, encore faut-il pouvoir la capturer dans le texte et l'identifier comme telle. Pour ce faire BIT-Alpha propose l'utilisation des caractères du code Unicode, puisse que BIT-Alpha édite le texte codé en Unicode, c'est-à-dire avec 20 bit par



Police : Arial Unicode MS
Style : Standard
Taille : 10
Couleur : Noir

Il s'agit d'une police OpenType qui sera utilisée à la fois sur votre imprimante et votre écran.

The OCR-Engine recognized this pattern with the following confidence:

This pattern corresponds to
● this String-Value: n̄
○ this Special-Value *UNRECOGNIZED*
4E53E596-948A-4A94-9CD0-C0B0CE7E067F

You can create a new Special-Value and add it to the OCR library:

You may choose another font for this pattern or set the currently selected font as default.
Arial Unicode MS □ Bold □ Italic

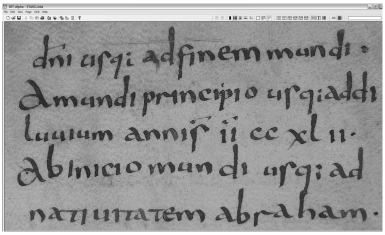
Press 'Update' to apply these modifications to the OCR library or press 'Delete' to delete the information that led to this result from the OCR library. If you do not desire to change the OCR library press 'Cancel'.

FIG. 9. Capture, apprentissage et reconnaissance d'une abréviation « n̄ ».

caractères. La figure 9 montre la capture d'une abréviation, retranscrite en symbole Unicode, (« n » avec barre supérieure d'abréviation).

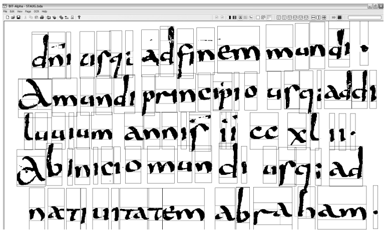
Ici, nous ferions référence à MUFI « Medieval Unicode Font Initiative ». Ce groupement de chercheurs propose l'utilisation et la normalisation des symboles Unicode pour transcrire les abréviations de textes manuscrits anciens. BIT-Alpha est en mesure, soit d'utiliser ces symboles Unicode pour les abréviations conformément à l'original du texte, soit d'en donner la transcription développée afin de rendre le texte directement lisible. Concernant le développement des abréviations, on pourra se référer aux travaux du Centre d'Etudes Supérieures de la Renaissance de Tours autour de Madame Marie-Luce Demonet et aussi à la publication sur « Les abréviatures » de la Renaissance (Andrieux-Reix et al.). Bien sûr, avec BIT-Alpha, tout opérateur reste entièrement libre de ses choix concernant la transcription et l'interprétation des abréviations.

Enfin, si BIT-Alpha peut faire la transcription, la corrélation, entre l'image scannée d'un texte et son édition en un fichier « texte-mode » utilisable informatiquement, il est



dni usq; adfinem mundi ·
Amundi principio usq; addi
luuium annis ii cc xl ii ·
Ab inicio mundi usq; ad
natiuitatem abraham ·

Fig. 10. Le document pris en charge dans BIT-Alpha, image 300 dpi couleur. Wolfenbüttel, Herzog August Bibliothek, Cod. Guelf. 67.5 Aug. 8°, fol. 3r



dni usq; adfinem mundi ·
Amundi principio usq; addi
luuium annis ii cc xl ii ·
Ab inicio mundi usq; ad
natiuitatem abraham ·

Fig. 11. L'image binarisée (Niblack modifié) et la capture des symboles.

aussi possible de faire la démarche inverse, c'est-à-dire d'apprendre, sur la base d'un texte connu, les caractères d'une écriture particulière, à partir du document scanné correspondant au texte connu. Cette démarche permet de valoriser des textes déjà transcrits par lecture humaine et saisis manuellement, en créant une bibliothèque de base de symboles ICR permettant de retranscrire ensuite d'autres textes scannés écrits de la même main. Si ce procédé d'apprentissage automatique de caractères est relativement facile à effectuer dans le cas d'une écriture où les caractères sont isolés ou bien formés, elle demandera l'intervention de l'opérateur dans le cas de caractères liés et imbriqués afin de bien définir l'endroit du tracé où la disjonction des symboles doit avoir lieu. Malgré tout, cela constitue une aide considérable à l'apprentissage ICR avec BIT-Alpha d'une écriture inconnue et par là de sa transcription.

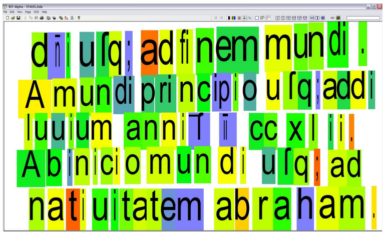
4.2 Edition du texte

Pour l'édition du texte, BIT-Alpha propose plusieurs possibilités.

Le texte pourra être édité soit en fichier au format XML contenant les mots et leur position, et aussi conformément à la norme ALTO ou TEI, soit sous forme d'un fichier au format PDF présentant le facsimilé du document comme image de fond avec le texte sous-jacent et invisible pour permettre un affichage, appelé « highlighting », en surbrillance dans l'original du texte du mot trouvé lors d'une recherche contextuelle. Un document manuscrit entièrement traité avec BIT-Alpha va maintenant être donné en exemple avec les figures 10 à 17.

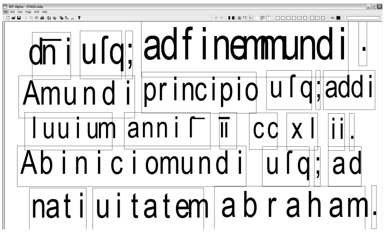
4.3 Identification du scribe

En paléographie, l'identification de l'auteur de l'écriture, respectivement de la personne l'ayant écrite de sa propre main, appelée ici scribe, est un sujet de grand intérêt.



dn̄i usq; ad finem mundi .
A mundi principio usq;addi
luuium anni ſ ii cc xl ii.
Ab inicio mundi usq; ad
natiuitatem abraham.

FIG. 12. La reconnaissance ICR des caractères, code Unicode (la couleur indique la fiabilité de lecture, du bleu « parfait » au rouge « douteux »).



dn̄i usq; adfinemmundi ·
Amundi principio usq;addi
luuium anni ſ ii cc xl ii.
Abiniciomundi usq; ad
natiuitatem abraham.

FIG. 13. Le texte transcrit (code Unicode) avec la formation des mots.

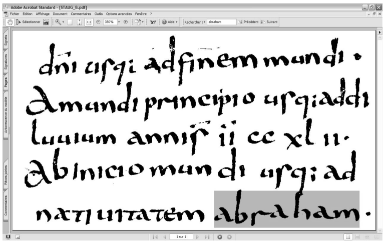


FIG. 14. Image d'un document PDF avec l'image binarisée en premier plan et le texte sous-jacent : mot recherché « abraham », affiché avec highlighting.

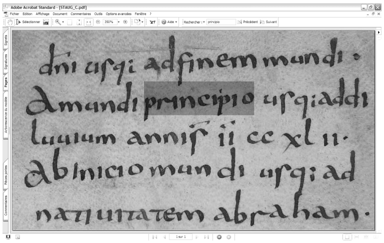
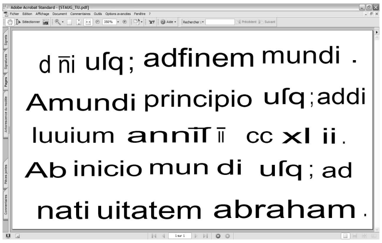


FIG. 15. Image d'un document PDF avec l'image couleur en premier plan et le texte sous-jacent : mot recherché « principio », affiché avec « highlighting ».

En partant de l'analyse détaillée des aspects géométriques du texte, BIT-Alpha peut être un outil d'aide à l'identification du scribe. Pour ce procédé plusieurs grandeurs analysées par BIT-Alpha sont disponibles :

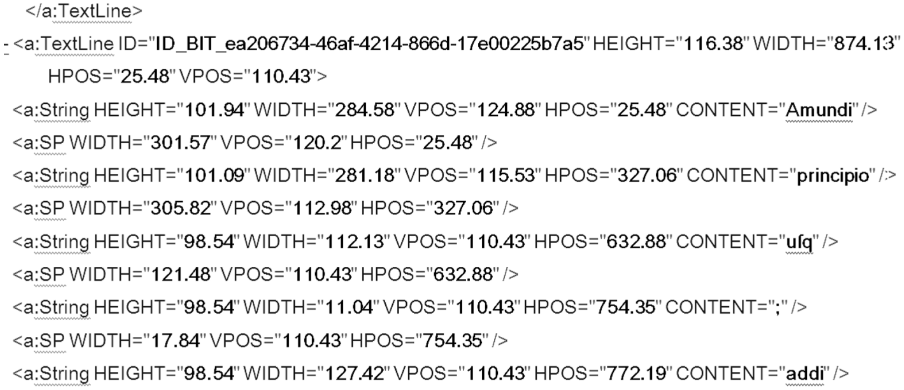
En premier, des grandeurs liées aux lignes :

- Epaisseur de ligne « e »: comme nous l'avons vue au chapitre 2, nous disposons de la distance entre base-ligne et top-ligne (voir figure 2).
- Hauteur de ligne : BIT-Alpha détermine en plus de la base-ligne et de la top-ligne, la ligne du haut des caractères appelée high-line et la ligne du bas appelée bottom-line. La distance entre bottom-line et high-line est aussi une caractéristique de l'écriture. Elle peut être exprimée relative à l'épaisseur de ligne « e ».
- Distance relative entre ligne : cette distance entre les lignes médianes à top- et base-ligne de lignes consécutives, sera exprimée relative par rapport à l'épaisseur



d n̄i uſq ; adfinem mundi .
Amundi principio uſq ;addi
luuium annīī ī cc xl ii .
Ab inicio mun di uſq ; ad
nati uitatem abraham .

FIG. 16. Image d'un document PDF avec le texte transcrit seul (Unicode).



```
   </a:TextLine>
- <a:TextLine ID="ID_BIT_ea206734-46af-4214-866d-17e00225b7a5" HEIGHT="116.38" WIDTH="874.13"
     HPOS="25.48" VPOS="110.43">
  <a:String HEIGHT="101.94" WIDTH="284.58" VPOS="124.88" HPOS="25.48" CONTENT="Amundi" />
  <a:SP WIDTH="301.57" VPOS="120.2" HPOS="25.48" />
  <a:String HEIGHT="101.09" WIDTH="281.18" VPOS="115.53" HPOS="327.06" CONTENT="principio" />
  <a:SP WIDTH="305.82" VPOS="112.98" HPOS="327.06" />
  <a:String HEIGHT="98.54" WIDTH="112.13" VPOS="110.43" HPOS="632.88" CONTENT="uſq" />
  <a:SP WIDTH="121.48" VPOS="110.43" HPOS="632.88" />
  <a:String HEIGHT="98.54" WIDTH="11.04" VPOS="110.43" HPOS="754.35" CONTENT=";" />
  <a:SP WIDTH="17.84" VPOS="110.43" HPOS="754.35" />
  <a:String HEIGHT="98.54" WIDTH="127.42" VPOS="110.43" HPOS="772.19" CONTENT="addi" />
```

FIG. 17. Image d'un extrait du fichier XML au format ALTO, donnant la position de chaque ligne et de chaque mot.

moyenne « e » des lignes. Cette grandeur exprime la place occupée par le corps de l'écriture par rapport à l'interligne.

- Imbrication du texte : la distance entre la bottom-line et la high-line de la ligne en dessous donne une idée de l'imbrication du texte, cette distance peut être négative et aussi mesurée relativement à la hauteur de ligne « e ».
- Espacement des mots : une distance moyenne entre mots, relativement à l'épaisseur « e » du texte, peut être mesurée par BIT-Alpha.
- Longueur des mots : à partir de la forme des mots identifiés, il est possible de donner une répartition statistique (répartition de Gauss) de la forme des mots, de leur longueur, par rapport d'une part à l'épaisseur « e » du texte et d'autre part par rapport à la hauteur des lignes et d'avoir aussi deux grandeurs aussi caractéristiques de l'écriture.

- Hauteur relative des mots : il est possible de donner la répartition statistique de la hauteur totale relative des mots, par rapport à l'épaisseur « e » du corps du texte. Ceci donne une grandeur significative des allongements de l'écriture, comme boucle de « l » et pied de « p ».
- Courbure des lignes : BIT-Alpha est en mesure de suivre la courbure d'une ligne de texte et ainsi de donner une valeur caractéristique de sa courbure relative. Ceci bien sûr n'a d'intérêt majeur que pour un texte écrit de main libre, sans trait « guide » sur le document.
- Variation de l'inclinaison : il est possible d'obtenir une valeur caractérisant la variation d'inclinaison moyenne d'une ligne à l'autre. Ceci n'a de sens que pour un texte écrit de main libre, sans trait guide sur le document.

La figure 18 montre en plus de la top-line et de la base-line, la bottom- line et la high-line pour chaque ligne du texte manuscrit. Par ailleurs les éléments de mesure de chaque élément graphique constitutif du texte sont visibles.

Nous avons en second les valeurs liées aux caractères eux-mêmes :

Nous pouvons extraire entre autres les paramètres suivants :

- Taille moyenne des caractères du texte.
- Taille moyenne des majuscules, des minuscules.
- Variance entre la plus grande et la plus petite lettre d'une ligne.
- Taille des allongements, relatifs à l'épaisseur du texte « e », comme des barres de « d », « l », « b » etc.
- Inclinaison moyenne des traits verticaux.
- Variance de l'inclinaison des traits verticaux.
- Taille moyenne des accents et points sur « i ».
- Distance moyenne entre les points et accents et la base-line.

Cette énumération n'est donnée qu'à titre d'exemple et il est certainement possible d'imaginer d'autres paramètres géométriques.

La détermination de ces éléments géométriques partant du texte dans son ensemble, des mots globalement et des caractères individuels, permettent de concevoir une matrice de paramètres caractérisant le texte et l'écriture de son scribe. En vue d'une comparaison de plusieurs textes pour en déterminer le scribe, ces paramètres peuvent aussi se voir attribuer une pondération donnant plus d'importance à une considération qu'à une autre. Il pourrait être ainsi possible de définir une « distance » entre écritures, qu'il serait possible de normer et définir ainsi une valeur « analytique » impartiale d'identification d'écriture, respectivement du scribe.[1]

Ainsi BIT-Alpha pourrait être un outil permettant de donner à l'identification du scribe, nous voulons dire de la main ayant écrit le texte, une approche « quantifiée »

[1] Cfr. la contribution de Aussems, Brink en cette volume.

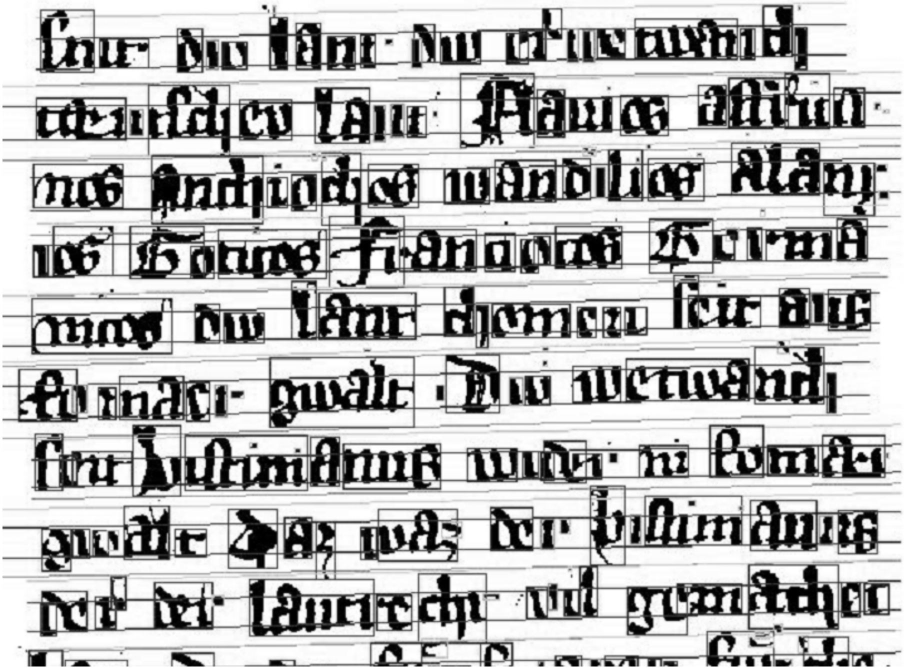
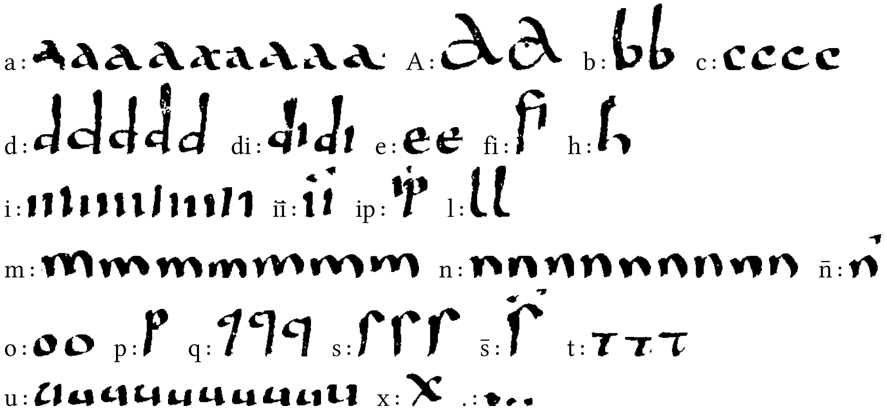


FIG. 18. Capture des éléments graphiques constitutifs d'un texte et des lignes bottom-line, base-line, top-line, high-line.

et « mesurée » pouvant aider le génie intuitif d'un paléographe. Ou permettre à une personne moins expérimentée de vérifier ses présomptions. Les idées, les possibilités et les considérations présentées ici devront être validées et affinées avec des chercheurs et paléographes. La validation de la détermination du scribe présumé d'un texte avec le logiciel BIT-Alpha_ICR passe aussi essentiellement pour la validation de la reconnaissance ICR des caractères du texte en utilisant une bibliothèque base de données ICR comportant les « empreintes numériques » des caractères du scribe considérés et préalablement appris.

Ainsi BIT-Alpha permet la reconnaissance de caractères par rapport à une bibliothèque base de données de symboles ICR pouvant être constituée séparément pour chaque scribe et appris directement d'un document traité. Chaque bibliothèque base de symboles ICR peut être prise en charge pour la reconnaissance au choix de l'opérateur. Ainsi en vérifiant la reconnaissance des symboles avec une base ICR d'un auteur connu, tout doute quant à la validation de l'auteur devrait pouvoir être levé. Le taux de fiabilité de lecture est directement affiché dans BIT-Alpha par la couleur des caractères proposés pour la transcription. La base ICR d'un scribe ne correspondant pas au tracé de l'écriture analysée donnera beaucoup de rouge dans la reconnaissance ICR.

Par ailleurs, BIT-Alpha permet de capturer et d'éditer en un clic tous les symboles graphiques d'une page traitée, en classant les lettres dans des dossiers distincts, un dossier pour chaque symbole reconnu. Les symboles capturés sous édités individuellement sous forme d'une image au format Bitmap et correspondant exactement au caractère capturé depuis le graphisme du texte. Nous les avons rassemblés en figure 18, chaque symbole étant une image individuelle utilisable à toute fin voulue.



Une étude de ces symboles devrait permettre au paléographe d'obtenir de précieux renseignements sur l'écriture du scribe et de bien pouvoir finaliser son identification. Il serait ainsi possible de générer une fonte de caractères pour texte permettant informatiquement d'écrire en reprenant les caractères du document.

5 Conclusion

L'approche informatisée du document manuscrit permet de prendre conscience des difficultés du traitement et de la transcription de ce type de document. Seule une approche prenant en compte l'ensemble des considérations, d'une part graphiques venant du tracé de l'écriture et d'autre part linguistiques venant de la langue utilisée, permettent un résultat utilisable.

Nous avons ici présenté nos développements et le logiciel BIT-Alpha, avec ses possibilités. Ce logiciel ICR peut constituer un outil d'aide à la transcription d'un texte manuscrit ou un outil d'aide à l'identification du scribe. Dans l'état de la technologie actuelle, nous considérons que l'intervention du spécialiste pour valider les résultats proposés reste essentielle. Cet exposé permet aussi de percevoir les difficultés de la transcription informatisée « pleine page » d'un texte manuscrit inconnu. Il reste encore un long chemin de recherche et développement avant une transcription automatique et nous ne voulons ici qu'apporter une contribution dans cette voie.

Mais malgré les difficultés, l'approche informatisée du document manuscrit est très prometteuse et la lecture ICR automatisée de textes manuscrits anciens ou en écritures liées, comme l'arabe ou l'hindi, permettrait l'accès de l'humanité à des trésors encore inaccessibles.

En paléographie, la définition d'une grandeur quantifiée mesurant impartialement la « distance » entre une écriture connue et une écriture analysée, apporterait une contribution d'intérêt aux débats concernant l'identification de certains écrits.

Bibliographie

ALTO : Analyzed Layout and Text Object. <http://www.ccs-gmbh.com/alto/>.

Andrieux-Reix, Nelly, Sonia Branca-Rosoff, and Christian Puech. « Les abreviatures à la Renaissance : enjeux et usages. » *Écritures abrégées (notes, notules, messages, codes).* Bibliothèque de Faits de Langues, org. Paris : Ophrys, 2004.

Emptoz, Hubert, Frank Le Bourgeois, Véronique Eglin, Stéphane Bres, Yann Leydier, Ikram Moalla, and Fadoua Drira. *Computer Assistance for Digital Libraries : Contributions to Middle-ages and Authors' Manuscripts Exploitation and Enrichment,* Second International Conference on Document Image Analysis for Libraries (DIAL 2006). Los Alamitos (CA) [et al.] : IEEE Computer Society, 2006. 265–80.

Le Bourgeois, Frank. *Automatic Metadata retrieval from Ancient Manuscripts*, Documents Analysis Systems (DAS 2004). Lecture Notes in Computer Science n° 3163. Berlin : Springer, 2004. 75–89.

Le Bourgeois, Frank, E. Trinh, Bénédicte Allier, Véronique Eglin, and Hubert Emptoz, *Document Image Analysis solutions for Digital libraries*, International Conference on Document Image Analysis for Libraries (DIAL 2004). Los Alamitos (CA) [et al.] : IEEE Computer Society, 2004. 2–24.

Le Bourgeois, Frank, Hubert Emptoz, Ikram Moalla, and Adel M. Alimi. *Contribution to the discrimination of the medieval manuscript texts : Application in the palaeography*, Document Analysis Systems (DAS 2006). Lecture Notes in Computer Science n° 3872. Berlin : Springer, 2006. 25–37.

Medieval Unicode Font Initiative. <http://www.mufi.info/>

MUFI character recommendation v. 2.0. Ed. Odd Einar Haugen, Bergen, 2006. <http://hdl.handle.net/1956/2003>.

Niblack, Wayne. *An Introduction to Digital image processing.* Upper Saddle River (NJ) : Prentice Hall, 1986. 115–16.

Ringlstetter, Christoph. *OCR-Korrektur und Bestimmung von Levenshtein-Gewichten.* Magisterarbeit Computerlingustik, Centrum für Information und Sprache, Ludwig-Maximilians-Universität-München, Prof. Klaus Uwe-Schulz, 2003.

The Palaeographical Method Under the Light of a Digital Approach

Arianna Ciula

Abstract

This paper has the twofold aim of reflecting upon a humanities computing approach to palaeography, and of making such reflections—together with its related experimental results—fruitful at the implementation level. Firstly, the paper explores the methodological issues related to the use of a digital tool to support the palaeographical analysis of medieval handwriting. It claims that humanities computing methods can assist in making explicit those processes of the palaeographical research that encompass detailed analyses, in particular of the handwriting and, more generally, of other idiosyncratic features of written cultural artefacts. Thus, palaeographical tools are to be contextualised and used within a broader methodological framework where their role is to mediate the vision, the comparison, the representation, the analysis and the interpretation of these objects. Secondly, the paper attempts to evaluate the experimentations carried out with a specific software and, in so doing, to test a humanities computing approach to palaeography at a practical level, so as to direct future implementations. Some of these implementations have already been carried out by the current developers of the application in question with whom the author collaborates closely, while others are still in progress and in need of future iterative refinements.

Zusammenfassung

Der Beitrag verfolgt ein doppeltes Ziel: Einerseits will er den fachinformatischen Zugang zur Paläographie allgemein reflektieren und andererseits, zusammen mit der Vorstellung von Ergebnissen einschlägiger Experimente, diesen Zugang in konkreten Anwendungen fruchtbar machen. Als erstes untersucht der Beitrag deshalb den Nutzen digitaler Werkzeuge zur Unterstützung der paläographischen Analyse mittelalterlichen Schreibens. Er kommt zu dem Ergebnis, dass fachinformatische Methoden dabei helfen, genau jene Prozesse paläographischer Forschung explizit zu machen, die Detailanalysen einschließen. Dies umfasst insbesondere die Analyse der Handschrift oder, allgemeiner, die Analyse von einmaligen Merkmalen schriftlicher kultureller Artefakte. So können paläographische Werkzeuge kontextualisiert und in einem weiteren methodischen Framework genutzt werden, wo sie eine Vermittlerrolle zwischen dem Aussehen, dem Vergleich, der Wiedergabe, der Analyse und der Interpretation dieser Objek-

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 219–235.

te übernehmen. Zweitens versucht der Beitrag Experimente zu bewerten, die mit einer speziellen Software durchgeführt wurden. Dabei wird der fachinformatische Zugang zur Paläographie auf einer praktischen Ebene erprobt und auf zukünftige Implementierungen hingearbeitet. Einige der Implementierungen wurden bereits von den Entwicklern in enger Zusammenarbeit mit der Autorin des Beitrags realisiert, während sich andere noch in Arbeit befinden und weiterer kontinuierlicher Verfeinerungen bedürfen.

1 Written Cultural Heritage and Trans-Disciplinarity

Generally described as the study of ancient writing devoted to deciphering and interpreting historical manuscripts and writing systems, palaeography has its most evident application in the process of identifying date and provenance of a particular script. A task that may seem rather circumscribed if it wasn't for its object of analysis—an old manuscript, be it a fragment, a whole codex, a roll or just a line of script running on the spine of an old book—which introduces substantial factors of complexity to the case. If digital technologies are to assist palaeographers, reflections on the complexity of the cultural artefacts under study are therefore indispensable. The identification of possible critical processes within the palaeographical method is also crucial.

Palaeography is by no means the only protagonist on the stage of disciplines that study the written heritage through its cultural artefacts; as stated by Julian Brown the scene is much richer:

> Palaeography means, in the strict sense, the study of ancient handwriting, and its basic objects are these: first, to read ancient texts with accuracy; secondly, to date and localize their handwriting. [...] The questions that palaeographers try to answer about a book are these. How, when, where, by whom and for whom was it first made? How has it been altered since? Who have owned it and used it? [...] You will understand that a palaeographer has to do his work on script and books with one hand. The fingers of the other must all be reserved for putting into a wide and appetizing range of different pies, from philology to the history of art. (Bately, Brown, and Roberts 17)

In less metaphorical terms, palaeography cannot proceed without sharing methods, tools and outcomes with co-disciplines such as epigraphy, codicology, philology, textual criticism—to name but a few (see Figure 1 for an attempt at representing these disciplinary clusters).[1]

[1] It is interesting to note that, in the past, palaeography has struggled to be recognised as a discipline against the conviction that its role was rather the one of "handmaid", simple instrument of, in turn, history, philology, literature, art history, archaeology, epigraphy, and diplomatics. Quoting Julian Brown once

Thus, inter-disciplinarity or trans-disciplinarity is a framework that it is not possible to prescind from, when it deals with the design of a tool or a set of tools to support the analysis and interpretation of a written object.

Indeed, according to Boyle (xv) 'integral palaeography' is the study of the script as intimately related to the history of the object that bears it. Therefore, besides the script itself, a manuscript is studied through other clues to its making, its functions and uses, its philological sources and textual tradition, its provenance and biography. The clues are multiple: the development of the various crafts involved in the manufacture, the notes made by scribes or illuminators, the indications borne, for instance, by liturgical texts, such as kalendars and litanies, the history of provenance and textual tradition, the factors of decoration; all the above are valuable guides to be variously interpreted.[2]

Therefore, independently from its more or less limited scope, the more any digital tool or resource—being it a digital facsimile of a manuscript, an application to segment letter forms, a digital edition, or an electronic publication of other kind—can be integrated within an environment where complementary material is also accessible, the more it becomes exponentially useful to the palaeographer.

Moreover, if we agree with Ginzburg that the humanities in general deal with "minute investigations of even trifling matters, to discover the traces of events that could not be directly experienced by the observer" (1989, 103), a tool that supports palaeographic research and its conjectures should make explicit those processes of the palaeographical method which apply to detailed analyses of individual entities, so as to facilitate broader intellectual operations (or scholarly primitives, as described by Unsworth) involved in investigations of this sort: analyses, comparisons, and classifications.

2 Quantitative and Digital Palaeography

Despite being very much debated in the history of the discipline as non-orthodox methodologies,[3] statistical and mathematical approaches have been applied in palaeography in the past.[4] More recent sporadic attempts have been made to develop or adapt computational tools to support palaeographical analysis.[5] However, compared to the

more: "Palaeographers, like scribes, were useful; [...] but not much was expected of them, and if they contributed to the progress of history and philology, it was only as the tools of better men." (Bately, Brown, and Roberts 17).

[2] Indeed, palaeographers treasure any visual representation of the manuscript sources under study and are eager to see more comprehensive image collections of such material made available and freely accessible in electronic form. For a recent discussion on the utility of digital resources for palaeographers see Dutschke.

[3] See in particular Poulle, D'Haenens, Ornato, Costamagna et al.

[4] See, for instance, Gilissen, Colloques Internationaux du CNRS, and Gumbert.

[5] See McGillivray, Moalla et al., Terras and Robertson, Terras, Stokes, and the following digital resources: CEEC, CDFP, EPPT, MANCASS C11.

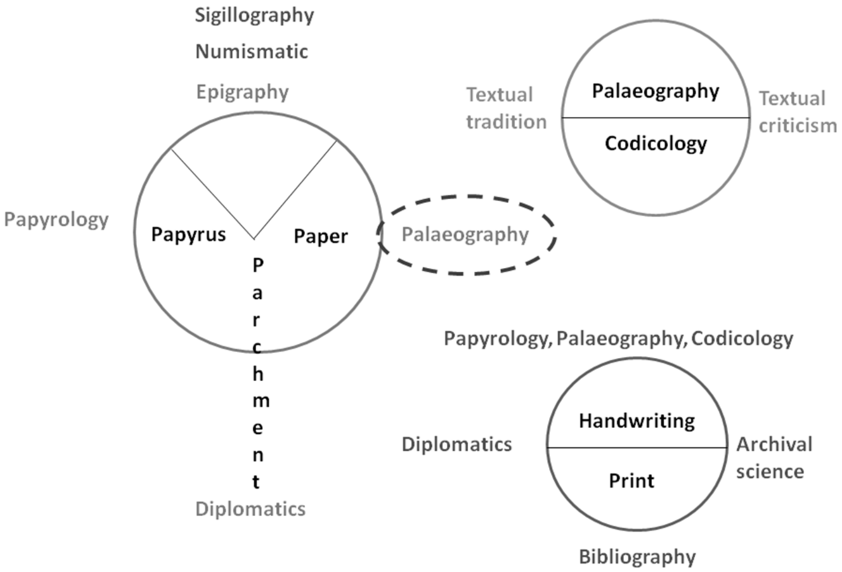


Figure 1. Disciplines that study the written cultural heritage (©Agati; note that the original chart was coloured, translated from Italian to English and slightly modified by the author).

state-of-the-art, the approach presented here is still characterised by a certain novelty and although in need of numerous improvements, has contributed to push forward the current developments of the computing application SPI (System for Palaeographic Inspection), the first design of which dates back to the 1990s.

The software in question (see Figure 2 for a drawing of its architecture and the interaction between its modules)—developed at the University of Pisa (Aiolli et al.) and currently being updated and improved at the University of Padova[6]—and its application to a specific corpus were described and discussed by the author in previous publications (Ciula 2003-2005).[7]

[6] The team of students in Computer Sciences at the University of Padova (Italy) includes Marco Dal Monte and is supervised by Professor Fabio Aiolli.

[7] The computing components of SPI were carefully tested and subjected to technical evaluation by its developers. However, the application had not beeen tested on a palaeographical corpus before the research project conducted by the author and summarised here and elsewhere (Ciula 2003-2005). The experimental use of SPI on a set of manuscripts was carried out as part of the author's PhD thesis (Ciula "Paleografia e

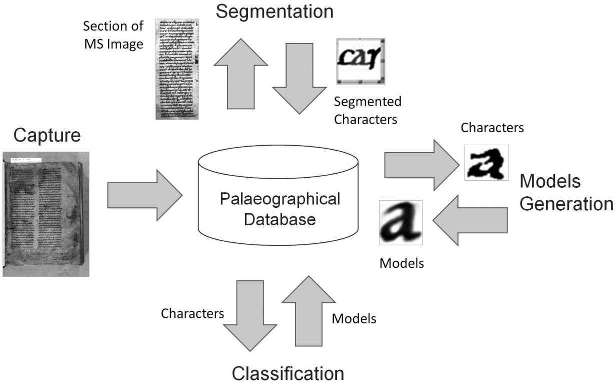


Figure 2. Architecture of Software for Palaeographic Inspections (SPI) developed at the University of Pisa (originally published in Ciula "Digital palaeography: using the digital representation of medieval script to support palaeographic analysis").

However, it is necessary at least to summarize briefly this research project which consisted in the following phases: (a) scanning of sample leaves of around forty codices held at the public library of Siena (*Biblioteca degli Intronati di Siena*); (b) image pre-processing for the insertion of the digital images in a relational database—which is the core component of the SPI software;[8] (c) segmentation of the relevant letters and ligatures (see Figure 3); (d) automatic generation of the letter models.

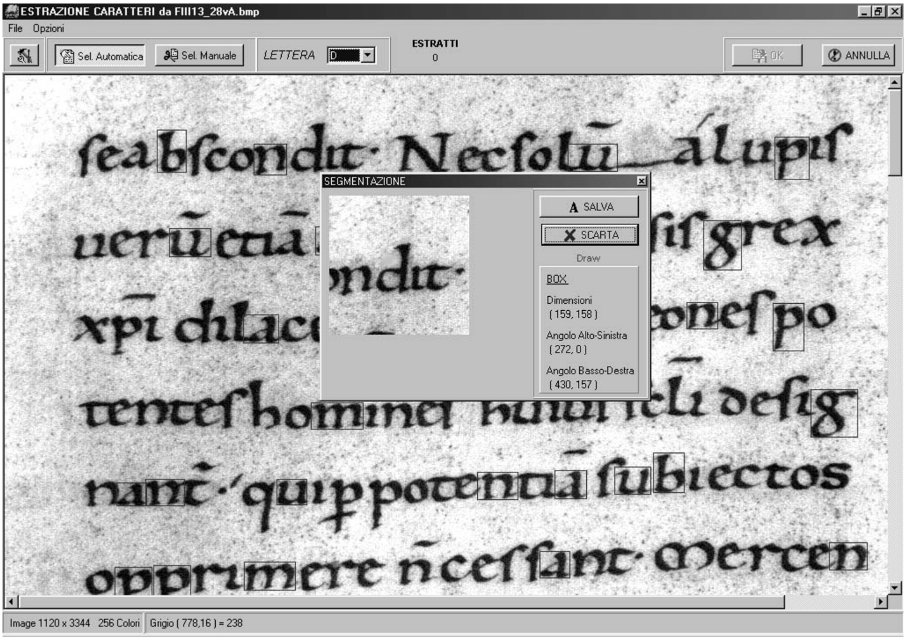
This approach of digital palaeography consists both of the process of preparing and collecting image data and, more innovatively, of the interpretation supported by the letter models (see Figure 4). The data collection was based on specific criteria to guarantee consistency across the chosen set of manuscripts. After the choice and definition of a palaeographical corpus took place, the following stages had to be thoroughly planned and documented: (1) definition of the digitisation criteria; (2) refinement and evaluation of the segmentation into letters and ligatures; (3) setting of the parameters for the letter model generation.

The recent but dense history of undertakings in manuscript digitisation and in image pre-processing for machine learning purposes,[9] especially the pattern recognition studies—some of which are at the base of the development of optical character recogni-

Informatica") on Manuscript and Book Studies (*Scienze del Libro*) at the University of Siena (Italy) completed in June 2005.

[8] The images were captured at a resolution of 300 dpi and archived in TIFF format before being converted to bitmap, cropped into sections corresponding to columns of handwriting when possible, and loaded into the application.

[9] For an introduction see Bunke and Wang.



LETTERA D

ESTRATTI 0

BOX

Dimensioni (159, 158)

Angolo Alto-Sinistra (272, 0)

Angolo Basso-Destra (430, 157)

Image 1120 x 3344 256 Colori

Grigio (778,16) = 238

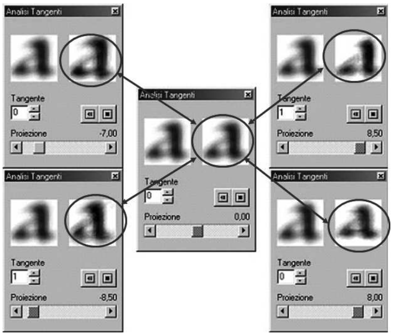
Figure 3. Example of segmented letter *d* within the SPI segmentation module.

tion systems known as OCR—served as background and supported the decisions while gathering palaeographical data.

On the other hand, the interpretative phase based on the analysis of the letter models and their automatic clustering has required insights into a much more established tradition of *doing* palaeography.[10] The comparison of types of letterforms—which is the main objective of analytical palaeography—has not effectively been supported so far by any tool. Therefore, the major challenge was represented by the attempt to integrate and support the palaeographical method within a digital humanities (as defined by McCarty 2003, 2005) research approach.

The experimental study has been carried out on a corpus of manuscripts written in different Caroline scripts, dating back from the tenth to the twelfth century and almost in every case certainly localisable in the area around the city of Siena in Tuscany or,

[10] On disquisitions around the palaeographical method see Costamagna, Pasquale, Ginzburg, Petrucci, and Supino Martini. See also Davis' paper on the supposed differences and commonalities between the palaeographical method and forensic document analysis.



Analisi Tangenti
Tangente 0
Proiezione -7,00

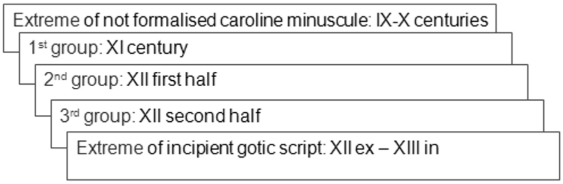
Analisi Tangenti
Tangente 1
Proiezione 8,50

Analisi Tangenti
Tangente 0
Proiezione 0,00

Analisi Tangenti
Tangente 1
Proiezione -8,50

Analisi Tangenti
Tangente 0
Proiezione 8,00

Figure 4. Example of model and dynamic graphical variations for the letter *a* as generated by SPI.



Extreme of not formalised caroline minuscule: IX-X centuries

1st group: XI century

2nd group: XII first half

3rd group: XII second half

Extreme of incipient gotic script: XII ex – XIII in

Figure 5. General palaeography of the corpus before the experimentation carried out with SPI.

more generally, in central Italy.[11] The direct results of the study—which consist mainly of a reclassification of the palaeographical corpus under examination as summarised in Figure 5 and 6—are thus based on the analysis of these specific occurrences of the Caroline minuscule and, eventually, on the regional evolution of the script. They provide a reorganisation of the corpus and of the script variants based on an integrated approach of computational and traditional palaeography.

2.1 Processes of Abstraction

However, the aim of this paper is to overtake the idiosyncratic interpretations related to this particular research project, so as to draw a wider picture. What are the methodological implications that arise from the unusual combination of analogical methods of letterforms description with the constraints and added values of a digital tool? It is worth noting that two main processes of abstraction had to be undertaken so as to make use of the abovementioned tool: firstly, the process of defining a taxonomy, a

[11] The initial categorisation is based on relevant catalogues and previous literature; in particular, see Avitabile et al., Garrison, Klange Addabbo, Cao et al.

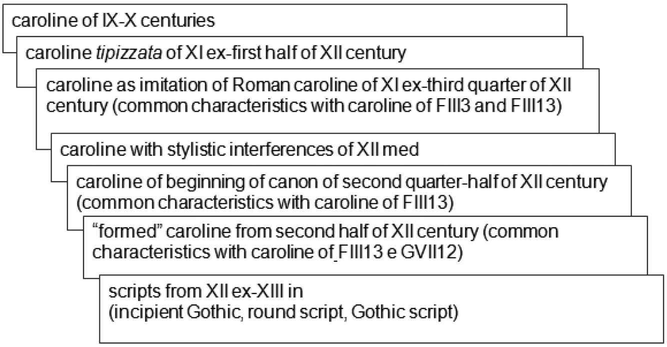

caroline of IX-X centuries
caroline *tipizzata* of XI ex-first half of XII century
caroline as imitation of Roman caroline of XI ex-third quarter of XII century (common characteristics with caroline of FIII3 and FIII13)
caroline with stylistic interferences of XII med
caroline of beginning of canon of second quarter-half of XII century (common characteristics with caroline of FIII13)
"formed" caroline from second half of XII century (common characteristics with caroline of FIII13 e GVII12)
scripts from XII ex-XIII in
(incipient Gothic, round script, Gothic script)

Figure 6. Re-organised palaeography of the corpus after the experimentation carried out with SPI (relevant manuscript signatures are used as references).

nomenclature to abstract from and *reduce* the 'polyphony' of individual manuscripts; secondly, the process of manipulating the digital images to abstract from and reduce the original morphology of specific letters.

The former process is certainly part of the traditional palaeographical method and was very much facilitated by the use of this tool, where every single graphical instance is decomposed into models that can be described one by one. However, beyond verbose narratives, a 'descriptive protocol' for such an annotation—comparable to the "analysis by chart method" used in forensic studies (Davis 258) —is not structured or configurable yet within the tool. On the contrary, the latter process consists mainly on a graphical approximation, and this is where the strength of the tool can be tested: SPI digests images and creates graphical digital letter models out of them, but, once more, not without technical limits as dealt with below.

2.2 Integral Palaeography

As stated above, the methodology behind this doctoral thesis was inspired by the concept of integral palaeography. Therefore, the corpus of manuscripts was analysed by concentrating both on the centrality of script—that is to say on the evolution of the Caroline minuscule in a localised region, on its relations with other styles, in particular on the influence of Roman book production—, as well as on the material culture aspects of the manuscripts, with the aim of integrating the digital models and the clues to their context and provenance as much as possible. In brief, this was achieved through the study of various characteristics of codicological and contextual nature in connection with the minute observation of the handwritten folia and of their mise en page. Indeed,

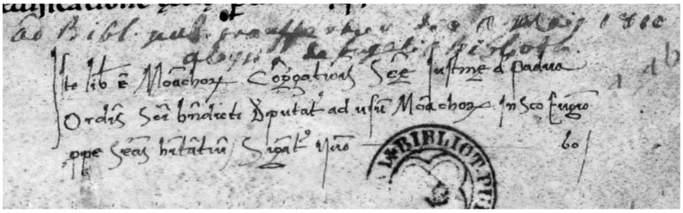


Figure 7. Codex FV21, Biblioteca Comunale degli Intronati (Siena, Italy), 1r. Three strata of information coexist here: the XV century note of possession (Monastery of S. Eugenio—Siena) in black ink, the XIX note of acquisition by the public library in red ink and the current stamp of the library.

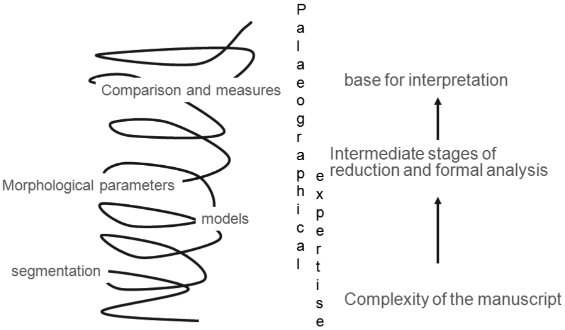


Figure 8. Representation of the intermediate stages of analysis leading towards the integration of the material history of the manuscripts and the digital models of its letter forms.

some of these observations have led towards the patrimonial contextualization of some of the codices (see Figure 7). It could be said that throughout the research process as supported by SPI, the complexity of the manuscript object was filtered through intermediate stages of reduction and formal analysis, so as to facilitate further deconstructions of the corpus eventually leading towards intepretative speculations (Figure 8). These intermediate stages of abstraction—namely, segmentation process, model generation, setting of morphological parameters, comparisons and measurements—are carried out while using the tool, but are not comprehensively and systematically supported by the tool itself.[12]

[12] As anticipated, the development of the tool in question was interrupted and has only recently been reconsidered and planned in collaboration with a team of computer scientists now based at the University of Padova (Italy).

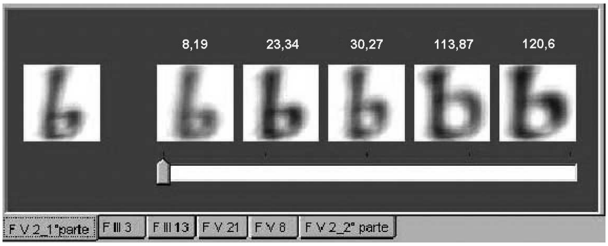


Figure 9. Example of the use of the diagram tool within SPI to compare various model of b in quantitative terms.

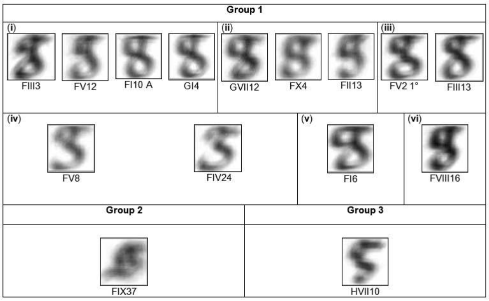


Figure 10. Categorisation of some digital models of the letter g.

2.3 Quantitative and Representational Value of the Models

The main powerful function of the SPI tool is its ability to compute graphical *features*. Indeed, the graphical models are digital in the sense that the morphological characteristic of the letter forms they encompass are expressed in quantitative terms. Thanks to the so called tangent-distance algorithm, the models can be compared numerically with the use of different tools internal to the system (the diagram in Figure 9 is one example of these).

Furthermore, the models bear a representative value. They incorporate the script or hand variants visually by making these variations perceivable to the eye. Therefore, the palaeographic analysis is forced to be anchored to the models as visible abstractions, *perceptible* prototypes (see Figure 10). The relativity of the research is then balanced by a somehow strengthened rigour in the method: the traditional formal-analytic approach is reinforced and modified by the use of a computing tool.

The question of the development of better tools is then: How much of the palaeographical expertise can the tool or its modules incorporate? If the use of the tool itself contributes to define, refine and enrich the underlying method, to what extent can this process be fed back into the tool and make it more sophisticated? In other terms, with respect to methodological issues, the room for improvement of both a new method of digital palaeography and a more sophisticated tool to perform it, lies in the gap between what is or can be formalised in line with the traditional palaeographical method and how the use of the tool forces to formalise further.

3 Overcome Limits and Future Perspectives

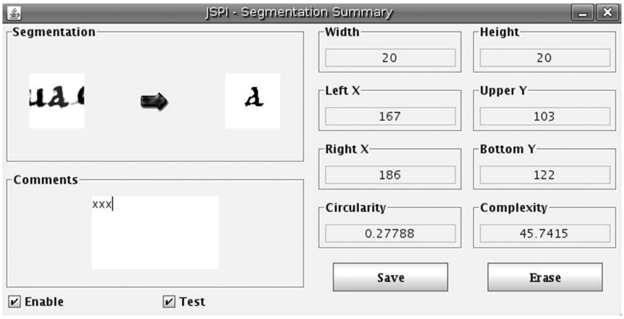
Such formalisations not only need to be technically robust and viable, but—in relation to what is stated above about the nature of palaeography as a cooperative discipline—they also need to account for some possible integration with other computational approaches to the study of manuscripts.

For instance, it would be desirable to be able to envisage a palaeographical module—consisting of functionalities such as the ones encompassed by SPI—within a wider and much more debated framework: the development of tools, including web services, for the creation and annotation of digital editions (Bozzi; Burnard, O'Brien O'Keeffe and Unsworth; Ciula and Stella; Iacob et al.; McCarty 2002; McGillivray 2006; Robinson).

To this end, it is necessary to make some more technical considerations regarding the SPI application and its recent developments. Indeed, besides the need for a methodology to be fairly documented, so as to be useful to other case studies, there are two sets of issues and challenges to consider: the ones already reported and dealt with in collaboration with the current developers of SPI in its bright new vest as JSPI (Java System for Palaeographic Inspections), and the ones still to be tackled. The following comments attempt to merge these and report on some of the overcome limits or first solutions to be refined further in connection with the issues still to be tackled for future development.

1. Documentation and transparency
   The alpha version of JSPI is written in Java, supported by MySQL and by other standard technologies which are well established, open-source and as such reliable and sustainable for the future. In addition, the release itself is accompanied by a descriptive handbook (written in Italian and currently being translated into English with the possibility of making it available in other languages once the application has been publicly released) which is both user as well as developer oriented.

   With respect to transparency, the interface of JSPI tends to be more informative than SPI; for instance, the window which visualizes the performed segmentation of a character also provides the values of the relevant features, such as coordinates and circularity, being measured on the pattern under study (see Figure 11), while the



JSPI - Segmentation Summary

Segmentation

Comments: xxx

Width: 20

Height: 20

Left X: 167

Upper Y: 103

Right X: 186

Bottom Y: 122

Circularity: 0.27788

Complexity: 45.7415

☑ Enable

☑ Test

Figure 11. JSPI window which visualizes the features being measured on the pattern being segmented.

dendrogram view has a mouse-over facility to visualise information on the patterns being examined.

2. Use of standards, extensibility and interoperability
In addition to what was mentioned under point (1) in relation to the use of standard programming languages and tools, it should be noted that the licenses of the whole set of technologies employed within JSPI allow free use.[13] This means that anybody could install the environment suitable for the software to work for free: a non-commercial and rather attractive solution for any palaeographer or humanities scholar.

Furthermore, JSPI is platform independent and therefore overcomes the main limit of its predecessor SPI, which could operate only within a Windows 98 platform.

In addition, a potential user-developer could modify the code of the JSPI software by downloading the Java Development Kit known as JDK and, by doing so, *extend* the functionalities of the application. Future developments will explore the possibility of incorporating the use of additional standards for the modelling of the data handled by the application (see point 3 below).

The use of MySQL to manage JSPI in its current alpha release was the strategy adopted to move towards the implementation of a full web service in the future. Indeed, the choice of MySQL as DBMS (Database Management System) implies already the concrete potential of operating in a networked environment, where the relational database could live in a remote server. This means that JSPI users could access this centralised database within a network, and in doing so, share the same data, for instance, by visualising the same manuscript images, letter models, and diagrams.

[13] This applies to the IDE (Netbeans) as well as to MySQL and the libraries (Java, Jama, and MySQL JDBC connector).

3. Refinement of SPI functionalities
   a. Image pre-processing: better filters for image pre-processing are required to overcome difficult segmentation cases when the manuscript handwriting is damaged or particularly complex to isolate;
   b. Image processing: as its predecessor, JSPI accepts only bitmap of 24 RGB values at the resolution of 300 dpi, while more flexibility or automatic conversion procedures could be implemented;
   c. Segmentation: the grid from where to select letters and ligatures for segmentation is extensible within JSPI, so that a user can define and extend as appropriate the 'alphabet' or set of letterforms according to the style or hands under study; moreover, besides the possibility of performing the segmentation manually, which was an option also available within SPI, the choice between multiple segmentations is offered to the user.
   d. Textual description
      i. Descriptive encoding: some fields within the application could be refined to allow for complex expressions, for instance of date, and descriptions possibly to be exported by using a standard such as the Guidelines of the Text Encoding Initiative (TEI Consortium);
      ii. Connection between images and text: again, the possibility to export the association between descriptions of specific palaeographical properties and the coordinates within a manuscript image in a standard format such as the encoding proposed by the TEI facsimile module or SVG (Scalable Vector Graphics) would be a step towards dealing with this challenge;
      iii. Search functionalities: in relations to the two points above, the search within the relational database which constitutes the backbone of JSPI could be structured and visualised, so as to allow for more sophisticated queries to be performed, saved and exported.
4. Graphical interface
   JSPI interface is now in English (SPI was in Italian) and it incorporates more elegant graphical solutions compared to those offered by SPI. It is, however still in need of further tests and improvements to become more usable and accessible.

4 Conclusions

In conclusion, despite various limitations, the specific research carried out with SPI on the palaeographical corpus from Siena assisted both in reorganising the corpus of manuscripts under study, in leading the work of computer scientists in improving the development and design of the application, and in reflecting on broader methodological issues.

If there are any successes to report in relation to this undertaking, they are mainly due to the benefits of a collaborative endeavour between the author as stubborn digital humanist practitioner and the computer scientists as long-sighted developers who, beyond a dusty and mysterious discipline, glimpsed a field in which it was worth investing. On the other hand, if there are failures besides the ineptitude of the author, these are to be attributed to the difficulty of maintaining a project, which was never formally funded, across countries and disciplines. It is just one of the kind of interdisciplinary projects which are needed in the humanities, but for which sustainable funding models are still lacking.

Bibliography

Agati, Maria Luisa. *Il libro manoscritto. Introduzione alla codicologia.* Studia archaeologica 124. Roma: L'Erma di Bretschneider, 2003.

Aiolli, Fabio, Maria Simi, Diego Sona, Alessandro Sperduti, Antonina Starita, and Gabriele Zaccagnini. "SPI: a System for Palaeographic Inspections." *AIIA Notizie* 12. 4 (1999): 34–38. <http://www.dsi.unifi.it/AIIA/ABSTRACT/paper405_99.html>.

Avitabile, Lidia, Maria C. Di Franco, and Viviana Jemolo. "Censimento dei codici dei secoli X-XII." *Studi medievali* 11.2 (1970): 1075–1101.

Bartoli Langeli, Attilio. "Ancora su paleografia e storia della scrittura: a proposito di un Convegno Perugino." *Scrittura e Civiltà* 2 (1978): 275–294.

Bately, Janet, Michelle P. Brown, Jane Roberts, eds. *A Palaeographer's View. The selected writings of Julian Brown.* London: Harvey Miller Publishers, 1993.

Boyle, Leonard E. *Medieval Latin palaeography: a bibliographical introduction.* Toronto: University of Toronto Press, 1984.

Bozzi, Andrea, ed. *Computer-aided recovery and analysis of damaged text documents.* Bologna: CLUEB, 2000.

Bunke, Horst and M.S.P. Wang, eds. *Handbook of character recognition and document image analysis.* Singapore: World Scientific Publishing Company, 1997.

Burnard, Lou, Katherine O'Brien O'Keeffe and John Unsworth, eds. *Electronic Textual Editing.* New York: Modern Language Association of America, 2006 (preprint). <http://www.tei-c.org/Activities/ETE/Preview/>.

Cao, Gian Mario, Tiziana Catallo, Mariella Curandai, E. Di Mattia, P. E. Fornaciari, E. Peruzzi, and F. Santi, comps. *Catalogo di manoscritti filosofici nelle biblioteche italiane.* Vol. 8. Firenze, L'Aquila, Livorno, Prato, Siena, Verona. Firenze: Olschki, 1996. 101–134.

Cuneiform Digital Forensic Project (CDP). University of Birmingham. <http://www.cuneiform.net>.

Codices Electronici Ecclesiae Coloniensis (CEEC), University of Köln. <http://www.ceec.uni-koeln.de>.

Ciula, Arianna. *A research project. The application of SPI Software to the Corpus of Manuscripts held in Siena.* MA thesis. King's College London, London, 2004.

Ciula, Arianna. "Digital palaeography." Poster presented at *Digital Resources for the Humanities* (DRH) conference. University of Newcastle, UK. 2004. <http://www.digitalmedievalist.org/journal/1.1/ciula/>.

Ciula, Arianna. "Modelli digitali di scrittura carolina." *Gazette du livre médiéval.* Autumn 45 (2004): 27–38.

Ciula, Arianna. "Digital palaeography: using the digital representation of medieval script to support palaeographic analysis." *Digital Medievalist* Spring 1 (2005). <http://www.digitalmedievalist.org/article.cfm?RecID=2>.

Ciula, Arianna. *Paleografia e Informatica. L'applicazione del software SPI al corpus di manoscritti senesi.* PhD thesis. Università degli Studi di Siena, 2005.

Ciula, Arianna. "L'applicazione del software SPI ai codici senesi." *Poesìa Medieval.* Ed. V. Valcàrcel Martìnez and C. Pérez Gonzàles. Colleccìon Beltenebros 12. Burgos: Fundaciòn Instituto Castellano y Leonés de la Lengua, 2005. 305–325.

Ciula, Arianna. "Zoom in, zoom out: la paleografia digitale tra sistema interdisciplinare e analisi dettagliate." *Griseldaonline* 6 (2007). <http://www.griseldaonline.it/informatica/6ciula.htm>.

Ciula, Arianna and Francesco Stella eds. *Digital Philology and Medieval Texts.* Pisa: Pacini editore, 2007.

Colloques Internationaux du CNRS, Les techniques de laboratoire dans l'étude des manuscrits, n° 548, (Parigi 13-15 settembre 1972) Paris, 1974.

Computers and the Humanities 36.1 and 36.3 (2002).

Costamagna, Giorgio. "Paleografia e Scienza." *Rassegna degli Archivi di Stato* 28 (1968): 293–315.

Costamagna, Giorgio, Leon Gilissen, Françoise Gasparri, and Alessandro Pratesi. "Commentare Bischoff." *Scrittura e Civiltà* 19 (1995): 321–352.

D'Haenens, Albert. "Pour une sémiologie paléographique et un histoire de l'écriture." *Scriptorium* 29 (1975): 175–198.

Davis, Tom. "The Practice of Handwriting Identification." *The Library* Sept. 8.3 (2007): 251–276.

Digital Image Archive of Medieval Music (DIAMM), University of Oxford and Royal Holloway University of London. <http://www.diamm.ac.uk>.

Dutschke, Consuelo W. "Digital Scriptorium: Ten Years Young, and Working on Survival." *Storicamente* 4 (2008). <http://www.storicamente.org/02_tecnostoria/filologia_digitale/dutschke.html>.

Garrison, Edward B. *Studies in the history of mediaeval Italian painting.* 4 vols. Firenze: L' impronta, 1984.

Gilissen, Leon. L'expertise des écritures médiévales. Recherche d'une méthode avec application à un manuscrit du XIe siècle : le Lectionnaire de Lobbes, codex Bruxelliensis 18018. Ghent: E. Story-Scientia, 1973.

Ginzburg, Carlo. *Clues, Myths, and the Historical Method.* Trans. John and Anne Tedeschi. Baltimore: Johns Hopkins University Press, 1989.

Gumbert, J. Peter. "A proposal for a Cartesian nomenclature." *Essays presented to G.I. Lieftinck, IV: miniatures, scripts, collections.* (Litterae Textuales) Ed. J.P. Gumbert and M.J.M. De Haan. Amsterdam: A.L. Van Gendt, 1976. 45–52.

Iacob, Ionut Emil, Kevin Kiernan, and Alex Dekhtyar. "Edition Production Technology: an Eclipse-Based Platform for Building Image-Based Electronic Editions." *Proceedings of ACH/ALLC Conference 2005* [Victoria, CA] 2005.

Klange Addabbo, Bente. *Codici miniati della Biblioteca comunale degli Intronati di Siena.* Siena: Edisiena, 1987.

MANCASS C11 Database, University of Manchester.
<http://www.arts.manchester.ac.uk/mancass/C11database/>.

McCarty, Willard. "A Network with a Thousand Entrances: Commentary in an Electronic Age?." *The Classical Commentary: Histories, Practices, Theory.* Ed. Gibson and Kraus. Leiden: Brill, 2002. 359–402.

McCarty, Willard. "Humanities computing." *The Encyclopedia of Library and Information Science.* New York: Marcel Dekker, 2003.

McCarty, Willard. *Humanities Computing.* Basingstoke: Palgrave MacMillan, 2005.

McGillivray, Murray. "Statistical Analysis of Digital Paleographic Data: What Can It Tell Us?" *TEXT Technology* 14.1 (2005): 47–60.
<http://texttechnology.mcmaster.ca/pdf/vol14_1_05.pdf>.

McGillivray, Murray. "Digitizing Sir Gawain: Traditional Editorial Scholarship and the Electronic Medium in the Cotton Nero A.x. Project." *Mind Technologies: Humanities Computing and the Canadian Academic Community.* Ed. Raymond Siemens and David Moorman. Michigan State University Press, 2006. 33–45.

Moalla, Ikram, Frank Lebourgeois, Hubert Emptoz, Adel M. Alimi. "Contribution to the Discrimination of the Medieval Manuscript Texts: Application in the Palaeography." *Document Analysis Systems VII, 7th International Workshop, DAS 2006, Nelson, New Zealand, February 13-15, 2006, Proceedings.* Lecture Notes in Computer Science 3872. Ed. H. Bunke and L. A. Spitz. Springer, 2006. 25–37.

Ornato, Ezio. "Statistique et Paléographie: peut-on utiliser le rapport modulaire dans l'expertise des écritures médiévales?" *Scriptorium* 29 (1975): 198–234.

Pasquale, Giorgio. "Paleografia quale Scienza dello Spirito." *Pagine Stravaganti di un filologo* Vol. 1. Firenze: Sanson, 1968. 103–117.

Petrucci, Armando. "La scrittura descritta." *Scrittura e Civiltà* 15 (1991): 5–20.

Poulle, Emanuel. "Paléographie et Méthodologie. Vers l'analyse scientifique des écritures médiévales." *Biblioteque de l'Ecole des Chartes* 132 (1974): 101–110.

Robinson, Peter. "The One Text and the Many Texts." *Literary and Linguistic Computing* 15.1 Apr (2000): 5–14.

Robinson, Peter. "Current issues in making digital editions of medieval texts—or, do electronic scholarly editions have a future?" *Digital Medievalist* Spring 1 (2005). <http://www.digitalmedievalist.org/journal/1.1/robinson/>.

Stokes, Peter A. "Palaeography and Image-Processing: Some Solutions and Problems." *Digital Medievalist* 3 (2007). <http://www.digitalmedievalist.org/journal/3/stokes/>.

Supino Martini, Paola. "Sul metodo paleografico: formulazione di problemi per una discussione." *Scrittura e Civiltà* 19 (1995): 5–29.

TEI Consortium, ed. *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* 1.3.0. Last updated on February 1st 2009. <http://www.tei-c.org/Guidelines/P5/>

Terras, Melissa and Paul Robertson. "Downs and acrosses: textual markup on a stroke level." *Literary and Linguistic Computing* 19.3 (2004): 397–414.

Terras, Melissa. *Image to Interpretation: Intelligent Systems to Aid Historians in the Reading of the Vindolanda Texts.* Oxford Studies in Ancient Documents. Oxford: Oxford University Press, 2006.

Unsworth, John. *Scholarly primitives: what methods do humanities researchers have in common, and how might our tools reflect this?* King's College London, 13 May 2000. <http://www.iath.virginia.edu/~jmu2m/Kings.5-00/primitives.html>.

The Computer and the Classification of Script

Mark Stansbury

Abstract

In the 1970s Bernhard Bischoff famously predicted that, thanks to technology, palaeography was on the road to becoming an art of measurement. The journey down this road has not been smooth, however, for several reasons. Although the idea of measurement seems uncontroversial, E.A. Lowe's attempt to measure the number of manuscripts written in half-uncial script shows that the script names that lie at the heart of palaeographical descriptions pose an insuperable problem, whether to man or machine. The reasons for this unsatisfactory system lie in the historical development of the discipline from its invention in the late-17th century. From the first, the names of scripts were used to localise manuscripts in time and place, and the names palaeographers use today are the direct descendants of these early systems. In the mid-20th century palaeographers began to focus on a different way of looking at script by exploring the strokes used to create the letters (*ductus*). These two approaches have led to a discipline divided between Linnaeans who emphasize taxonomy and Darwinians who emphasize evolution. Most digital palaeography has focused on the first, while the second could offer a richer vein to mine.

Zusammenfassung

Bereits in den 1970er Jahren sagte Bernhard Bischoff voraus, dass wegen der technologischen Entwicklung die Paläographie auf dem Wege sei, eine Kunst des Messens zu werden. Der Weg dorthin verläuft jedoch aus unterschiedlichen Gründen nicht geradlinig. Obwohl die Idee des Messens unproblematisch erscheint, zeigt E.A. Lowes' Versuch die Zahl der Handschriften in Halbunzial-Schrift zu »messen«, vor welchem Problem Mensch und Maschine stehen: Es sind die »Schriftnamen«, die als Grundlage der paläographischen Beschreibung dienen. Der Grund hierfür ist aus der Entwicklung der Disziplin seit ihrer Einführung im späten 17. Jahrhundert zu erklären. Von Beginn an wurden Schriftnamen benutzt, um Manuskripte in Raum und Zeit zu lokalisieren, und die Bezeichnungen, die Paläographen heute benutzen, sind direkte Abkömmlinge dieser frühen Bemühungen. Seit Mitte des 20. Jahrhunderts begannen Paläographen, Schrift auf eine andere Art zu betrachten, indem sie die Strichführung (Duktus) zur Erzeugung von Buchstaben untersuchten. Diese beiden Ansätze haben zu einer Disziplin geführt, die sich aufspaltet zwischen den Linnéisten, die die Taxonomie betonen, und

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 237–249.

den Darwinisten, die die Evolution hervorheben. Digitale Paläographie hat sich bislang zumeist auf den ersten Ansatz konzentriert, während der zweite eine fruchtbarere Ader bietet, die es zu erschließen gilt.

1 Introduction

In a much discussed passage in the introduction to his *Paläographie des römischen Altertums und des abendländischen Mittelalters*, Bernhard Bischoff wrote that thanks to technical means palaeography was on the way from being an art of seeing and aesthetic empathy (*Kunst des Sehens und der Einfühlung*) to becoming an art of measurement (*Kunst des Messens*) (Bischoff 2004 19). In the thirty years since he wrote those words, the technical means to accomplish the measurement he predicted have certainly grown prodigiously, primarily in two areas: first, the production of images through digital photography and their reproduction and distribution through the World Wide Web; and second, the analysis of those images by computer. Yet much evidence suggests that palaeography generally has not moved far down the road Bischoff predicted for it in 1978. For example, of the general introductions to the field written after Bischoff—Gasparri, Derolez, and Frenz updating Foerster—only Derolez even acknowledges the existence of the road that Bischoff saw. And far from sharing Bischoff's vision, Derolez called his choice of words 'regretful' and foresaw a limited role for quantitative data in palaeography (7-8). Even for those who see themselves travelling toward the same goal, that of making palaeography a *Kunst des Messens* using technology, there are many roads that seem to lead there, as the contents of this volume shows. Although palaeography was born of technological developments, trying to incorporate recent technology seems to have encouraged a fundamental re-examination of the aims of palaeography. This paper aims to ask why that re-examination is necessary and to suggest a few ways forward.

2 Measuring and Manuscripts

To begin with, Bischoff's statement has a vaguely oracular quality that may partially explain the amount of scholarly discussion it has stimulated. Though the discussions have focused on Bischoff's view of the future, his view of palaeography in his day is almost as interesting as his prediction for the future. By choosing to describe palaeography as an art of seeing and of *Einfühlung*, Bischoff was echoing a phrase of Joachim Kirchner's (Derolez 2 n. 3) and associating palaeography with aesthetic theories of the late-19th and early-20th century. The German *Einfühlung* is a calque of the Greek ἐμπάθεια and was first used in connection with aesthetics in 1873 by Robert Vischer in his

Über das optische Formgefühl: Ein Beitrag zur Aesthetik (Koss 139), with the earliest citation of English 'empathy' in this sense published in 1904 according to the *Oxford English Dictionary.* The principle Vischer proposed is that viewing inanimate forms involves the unconscious projection of the viewer's own bodily form onto the viewed object so that the experience of the object becomes in some sense corporeal, hence the possibility of empathy with inanimate forms (Koss 140). Bischoff's choice of words is thus more resonant than the translations 'seeing and understanding' (Bischoff 1990 3) or 'seeing and comprehending' (Derolez 2). Rather than basing the palaeography of his day on these relatively vague terms, Bischoff chose to found it on a late-nineteenth century aesthetic understanding of non-corporeal forms through the activity of seeing. But for Bischoff palaeography was and remained an art. German *Kunst* has much the same semantic range as English 'art' and by it Bischoff must have meant a set of techniques roughly corresponding to the late-antique and medieval *ars*, as opposed to the systematic body of knowledge implied by *Wissenschaft.* Most significant, however, is the radical simplification his statement predicted for the future: through technology, no longer would palaeographers look at letterforms and use aesthetic empathy to understand them; instead, they would measure. And of course *Messen* has caused most discussion because it changes the role of the palaeographer and because it offers a central place to the technical means that seemed to be around the corner—all the while remaining vague about what those technical means were. Whatever they were, however, they no longer required that palaeographers look and try to achieve an empathetic understanding of forms.

Given the state of computer analysis when Bischoff wrote, as well as the lack of interest his own work shows for it, this prediction seems to be more an educated guess about the relatively bleak future than a specific vision of it. The statement remains significant, however, because of the use others have made of it, for Bischoff's *Kunst des Messens* has come to represent the future of palaeography. One reason for this is that his term 'measurement' involves the recording of results obtained by observation (as opposed to seeing). When we measure the length of an object we observe where its edges fall in relationship to the marks on a ruler, for example. It is possible to perform this action with more or less accuracy, but once there is agreement about what is to be measured, the process is straightforward. Such observation can be replaced by technical devices: instead of comparing the edge of an object with the marks on a ruler, the display of a digital caliper can simply produce the result. Technical devices can substitute for observation in this sense because they can collect the data they are designed to collect, but they cannot see in the same way that humans can and they certainly cannot experience the kind of aesthetic empathy that Bischoff saw as central to palaeography.

Whatever exact future Bischoff predicted, any art of measuring must consist in establishing magnitudes that can then be compared. The most elementary form of such a method applied to manuscripts consists in measuring and looking for patterns. We may, for example, measure the dimensions of manuscript pages, the dimensions of their writing areas, and the number and position of the ruling lines and then record the results in a database. We can then ask whether the manuscripts in the database fall into categories, whether a given manuscript is related in format to any of the manuscripts in the database, whether that relationship is significant, etc. The scholars in *Mise en page et mise en texte du livre manuscrit* (Martin and Vezin) have shown that this process can lead to interesting results, to take just one group of examples. Most palaeographers would consider such work to be codicological rather than palaeographical, however, since it deals not so much with script as with the supports for script. Still, similar quantitative methods have been applied to script, such as the study of Bozzolo and Ornato on ligatures cited with approval by Derolez (8). Similar approaches, Derolez argued, ought to be confined to ones that count and measure 'significant features of handwriting' and chart the results (8–9). In other words, Derolez argued for looking for significant features, measuring them, then analyzing the results. Even this, however, is easier said than done and the reasons for this difficulty point to one of the great roadblocks on the way to the palaeography of the future.

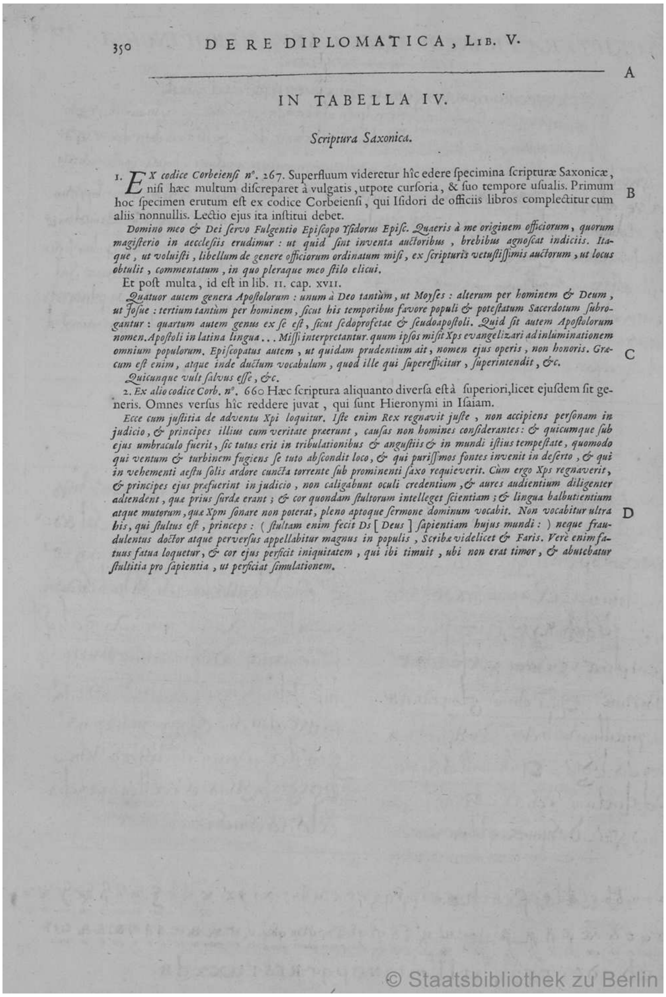
In order to demonstrate the problem, we can take E. A. Lowe's 'Hand-List of Half-Uncial Manuscripts', first published in the *Miscellanea Fr. Ehrle* in 1924. This project would seem to be exactly the sort of thing Derolez advised: Lowe examined manuscripts, 'measured' the ones in half-uncial script, and then gave the results in a list of 160 manuscripts. In even so elementary a task, however, there are difficulties, and Lowe meets them in characteristically cantankerous fashion. He is not interested, he writes, in futile discussions about whether half-uncial is a minuscule or a mixed script. 'The name "half-uncial" stands for a definite type and calls up a clear and distinct image' (Lowe 35). Yet in the next paragraph, the image turns out not to be so clear and distinct as one might hope, for in this paragraph Lowe laid out his rule for distinguishing half-uncial from uncial and minuscule. It would be easy if every scribe were like the scribe of the Basilican Hilary, he writes, but alas it is not so. Lowe called scripts like that of the Hilary manuscript 'canonical half-uncial', by which he means that the letterforms of these manuscripts closely match a single, idealized half-uncial alphabet, which is the rule (canon) that they obey. Unfortunately, such manuscripts represented only a portion of his corpus: '[…] there are many deviations from this norm, chiefly owing to the presence of uncial elements.' What's a palaeographer to do? Lowe, rarely at a loss, devised a rule: any script was half-uncial if it had four non-uncial letterforms, giving the examples of the letters **b**, **d**, **m**, **g**, **r**, **s**. Yet even then not everything fell into place: 'On the other hand, there are some eighth-century manuscripts, which show a curious mixture of uncial and half-uncial forms, which renders classification rather arbitrary' (Lowe 36). So not even the arbitrary rule designed to clarify an already clear image kept Lowe from using his judgment in the basic job of counting which manuscripts are written in half-uncial script. Ironically, then, Lowe's contortions serve to point out that 'half-uncial calls up a clear and distinct image' cannot be true. If this is so, what are we to make of claims that the name is 'useful and scientific'? What would happen, for example, if we changed Lowe's rule to three variant letterforms? It would not change the image of the script, but it would change the number of manuscripts in the list.

The purpose of this example is to show that statistical operations can be used in certain cases, i.e. those in which the classification of the things being enumerated is clear. Unfortunately, as palaeographers we often find ourselves in Lowe's position—discussing not how many half-uncial manuscripts there are but how to define a half-uncial manuscript. In other words, even if a machine existed that could automatically analyze scripts and match them to the idealized alphabets of the handbooks, thus classifying manuscripts automatically, it still could not solve the problem Lowe faced and that the palaeographers who followed him have, for the most part, avoided. In order to understand one way that technology can play a role in palaeography, then, we must first examine the naming of scripts.

3 Historical View of Script Classification

Technology and palaeography were intimately connected long before Bischoff wrote, of course. Indeed, palaeography was made possible by two technological innovations in the fifteenth century: printing and engraving. Engraving and later etching and lithography made possible the mass reproduction of identical facsimiles of scripts. These reproductions could then be compared with manuscripts and with each other, a comparison that would previously have required drawings or a good visual memory. Printing, then, made possible the mass reproduction of the analyses of the plates produced by engraving. It was one of Ludwig Traube's many insights that the history of palaeography continued to be linked with the techniques of reproduction, characterizing the discipline in his day as palaeography in the age of photography (Traube 57). It is thus an homage to Traube's work that we speak of palaeography as being in a digital age. The implications of this fact are that reproductions are more complete and more easily available than at any time in history.

As is well known, the first systematic classification of scripts grew out of the disagreement between the Maurist Jean Mabillon and the Bollandist Daniel Papebroch. In answering Papebroch's *Propylaeum antiquarium circa veri et falsi discrimen in vetustis membranis* published in 1675 in the *Acta Sanctorum*, Mabillon's *De re diplomatica* of 1681 formed the foundation of two disciplines: diplomatic and palaeography. The palaeographical part of his project grew out of Mabillon's wish to use script as one of the criteria for localizing a document in time and space. He summarized his position early in the work when he wrote that 'One way of writing was found amongst the Romans, another amongst other peoples [*nationes*]. One may identify almost as many



350 D E R E D I P L O M A T I C A , Lib. V. A

IN TABELLA IV.

Scriptura Saxonica.

1. *EX codice Corbeienſi n°.* 267. Superfluum videretur hîc edere ſpecimina ſcripturæ Saxonicæ, niſi hæc multum diſcreparet à vulgatis, utpote curſoria, & ſuo tempore uſualis. Primum hoc ſpecimen erutum eſt ex codice Corbeienſi, qui Iſidori de officiis libros complectitur cum aliis nonnullis. Lectio ejus ita inſtitui debet. B

Domino meo & Dei ſervo Fulgentio Epiſcopo Yſidorus Epiſc. Quaeris à me originem officiorum, quorum magiſterio in aecleſiis erudimur: ut quid ſint inventa auctoribus, brebibus agnoſcat indiciis. Itaque, ut voluiſti, libellum de genere officiorum ordinatum miſi, ex ſcripturis vetuſtiſſimis auctorum, ut locus obtulit, commentatum, in quo pleraque meo ſtilo elicui.

Et poſt multa, id eſt in lib. II. cap. XVII.

Quatuor autem genera Apoſtolorum: unum à Deo tantùm, ut Moyſes: alterum per hominem & Deum, ut Joſue: tertium tantùm per hominem, ſicut his temporibus favore populi & poteſtatum Sacerdotum ſubrogantur: quartum autem genus ex ſe eſt, ſicut ſedoprofetae & ſeudoapoſtoli. Quid ſit autem Apoſtolorum nomen. Apoſtoli in latina lingua . . . Miſſi interpretantur. quum ipſos miſit Xps evangelizari adinlùminationem omnium populorum. Epiſcopatus autem, ut quidam prudentium ait, nomen ejus operis, non honoris. Græcum eſt enim, atque inde ductum vocabulum, quod ille qui ſuperefficitur, ſuperintendit, &c. C

Quicunque vult ſalvus eſſe, &c.

2. *Ex alio codice Corb. n°.* 660 Hæc ſcriptura aliquanto diverſa eſt à ſuperiori, licet ejuſdem ſit generis. Omnes verſus hîc reddere juvat, qui ſunt Hieronymi in Iſaiam.

Ecce cum juſtitia de adventu Xpi loquitur. Iſte enim Rex regnavit juſte, non accipiens perſonam in judicio, & principes illius cum veritate præerunt, cauſas non homines conſiderantes: & quicumque ſub ejus umbraculo fuerit, ſic tutus erit in tribulationibus & anguſtiis & in mundi iſtius tempeſtate, quomodo qui ventum & turbinem fugiens ſe tuto abſcondit loco, & qui puriſſmos fontes invenit in deſerto, & qui in vehementi aeſtu ſolis ardore cuncta torrente ſub prominenti ſaxo requieverit. Cùm ergo Xps regnaverit, & principes ejus præfuerint in judicio, non caligabunt oculi credentium, & aures audientium diligenter adtendent, quæ prius ſurdæ erant; & cor quondam ſtultorum intelleget ſcientiam; & lingua balbutientium atque mutorum, quæ Xpm ſonare non poterat, pleno aptoque ſermone dominum vocabit. Non vocabitur ultra his, qui ſtultus eſt, princeps: (ſtultam enim fecit Ds [Deus] ſapientiam hujus mundi:) neque fraudulentus doctor atque perverſus appellabitur magnus in populis, Scribæ videlicet & Faris. Verè enim fatuus fatua loquetur, & cor ejus perficit iniquitatem, qui ibi timuit, ubi non erat timor, & abutebatur ſtultitia pro ſapientia, ut perficiat ſimulationem. D

Figure 1. Page 350 of the second edition of Jean Mabillon's *De re diplomatica* with commentary on Tabella IV, the script samples and alphabets on the facing page (Fig. 2). After giving brief information about the script and manuscripts Mabillon printed a transcription (sometimes abbreviated) of the facing passages. (From: Mabillon, Jean: *De re diplomatica libri VI – Ed. 2 ab ipso auctore recognita, emendata et aucta* – Paris 1709. In the Diez Collection of the Staatsbibliothek zu Berlin – Preußischer Kulturbesitz, Bibl. Diez fol. 615. Used with the kind permission of the owner and holder of the rights for the digital version, the Staatsbibliothek zu Berlin.)

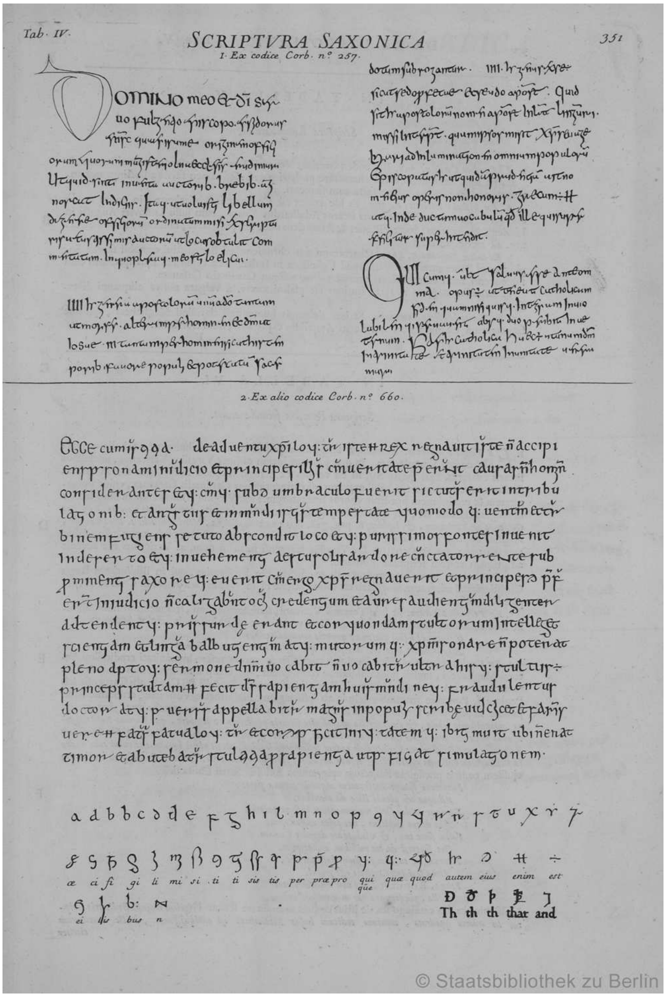

Tab. IV.
SCRIPTURA SAXONICA
351
1. Ex codice Corb. n° 257.
2. Ex alio codice Corb. n° 660.
æ ci fi gi li mi si ti ti sis tis per præ pro qui que quæ quod autem eius enim est
ei llis bus n
Th th th that and
© Staatsbibliothek zu Berlin

Figure 2. This page (351) faces the commentary printed in Fig. 1. The two engraved script samples are designed to show the reader a 'typical' manuscript in Mabillon's *scriptura saxonica.* Below the two script samples are alphabets intended to give an idealized version of the script. Note that an extra set of characters has been added with characters used to write the Anglo-Saxon language (the plate is unchanged from the first edition). (From: Mabillon, Jean: *De re diplomatica libri VI – Ed. 2 ab ipso auctore recognita, emendata et aucta* – Paris 1709. In the Diez Collection of the Staatsbibliothek zu Berlin – Preußischer Kulturbesitz, Bibl. Diez fol. 615. Used with the kind permission of the owner and holder of the rights for the digital version, the Staatsbibliothek zu Berlin.)

ways of writing as there are different peoples, though the ways of writing of each people is different at various times' (Mabillon 45). He went on to argue that as the Goths and Lombards conquered Italy, their scripts replaced Roman script. Because he linked script and national identity, Mabillon had no need to explain how the different national scripts were related; the only problem was to explain how the changes within the national scripts occurred over time so that these could be dated. Mabillon's presentation of his results in Book Five of *De re diplomatica* also set a pattern that palaeographical manuals follow to this day. We can examine Tabella IV and its facing page as one example. The right-hand page 351 (Fig. 2) shows examples of Mabillon's *scriptura saxonica* from two Corbie manuscripts (the manuscripts are now in St Petersburg, Lat. Q v I 15 and Lat. F v I 3 respectively Number One and Two), while the facing page (p. 350, Fig. 1) has a transcription and short commentary.

In presenting the scripts in this way in the plate Mabillon sought to give a sense of how texts looked on the page as well as a set of letterforms abstracted from the scripts of individual scribes. By juxtaposing the hand of a single scribe with an artificial alphabet, the page was designed to help someone classify a given manuscript (Mabillon 343). Eventually, these letterforms were treated like type, which is to say that they were considered as single graphic entities that arrived simultaneously on the written surface rather than being traced upon it in a series of strokes. The origin of this way of looking at script lies in the early attempts to create generalized alphabets that would represent a class of script—in other words something corresponding to the form the scribe had in his or her mind before writing. Once these idealized alphabets had been created, classification of the script found in a manuscript was a matter of finding the alphabet it most nearly resembled. This method has advantages in being able to localize scripts, but disadvantages in that it has no model for showing the relationships among scripts.

As influential as Mabillon's insight about the relationship between peoples and scripts was, it was soon challenged by Scipione Maffei, who claimed he could prove that such national scripts never existed in the same way he could prove a geometrical proposition (Maffei 113). Maffei argued that Mabillon's national scripts were simply modifications of three basic forms of Roman script: capital, minuscule, and cursive. In doing this Maffei argued for a change in the basis for Mabillon's system, but he did not argue against the structure erected upon it, i.e. the names for scripts or their dates. As important as Mabillon and Maffei were, it was the *Nouveau Traité de Diplomatique*, published between 1750 and 1765, that set palaeography on the trajectory it followed for some 200 years. Written by the Benedictines Toustain and Tassain, the work deliberately follows Mabillon and defends his approach against critics. In working out the implications of Mabillon's views, however, the *Nouveau Traité* did not reject Maffei out of hand, but remarked in the preface of the first volume that while Maffei's project claimed to speak more to the eye than to the intellect, their project attempted to speak both to the eye and the intellect (Toustain and Tassin 1 xxi). And in the next volume, the authors summarized Maffei's arguments against Mabillon, while only partially refuting them (Toustain and Tassin 2 283–88). And, of course, the *Nouveau Traité* introduced the term *demi-onciale*, which has bedeviled palaeographers to this day. So influential was this formulation that the writers of the *Dictionary of science, literature, and art* described diplomatic in 1867 this way: 'The principles laid down by Mabillon [...] were more fully developed about the middle of last century, in the *Nouveau Traité de Diplomatique,* which has left little to be done by subsequent labourers in this field beyond the duty of translation, compilation, or abridgement' (Brande and Cox s.v. diplomatic). But so it was not to be.

In 1937 Jean Mallon first addressed the problem that had lain dormant since Mabillon: how could letters evolve? He followed that article with a demonstration (both in print and on film) of how the letter **b** might have evolved and he ended the decade by producing along with Robert Marichal and Charles Perrat a volume of plates and transcriptions entitled *L'écriture latine de la capitale romaine a la minuscule.* Nothing, at first glance, could possibly seem less controversial; yet in the introduction there are rumblings of disquiet about the state of a discipline confined to the field allotted it in the seventeenth and eighteenth centuries and only timidly venturing out. The authors contented themselves with remarking that any justification for breaking down the barriers would be premature (Mallon et al. 2). A hint of what was to come, however, lay in the fact that the facsimiles were organized into three groups—scripts written on hard materials, the scripts of documents, and the scripts of books—and that none of the scripts had names. The promised *exposition de doctrine* came first in Marichal's lengthy 1948 article entitled *De la capitale romaine a la minuscule* (Marichal) and then in Mallon's *Paléographie romaine* of 1952. These publications not only argued that palaeographers ought to consider all the written monuments of a period but also gave a tool for analyzing scripts using the movements of the pen (*ductus*) to show the similarities underlying many seemingly diverse letterforms. As we saw with the volume of plates from 1939, this focus on continuity and evolution made the classification of scripts, much less their names, almost irrelevant. And although his critique of the naming system is devastating and he complained bitterly about the refusal of these inadequate and inaccurate terms simply to disappear (Mallon 1982 261), he never seems to have recognized the need they fill, much less offer an alternative. Perhaps predictably, then, many palaeographers writing in English and German tend not to adopt Mallon's approach to the subject or even at times to be aware of its existence as an alternative. As Julian Brown put it in 1982: 'Silly though they undoubtedly are, most of us have declined to accept Mallon's contention that the scripts should be renamed in some manner that reflects their natures and historical relationships to each other: for one thing, they are too deeply embedded in the literature, and for another we all know what they mean' (43). Palaeographers writing in French and Italian, on the other hand, have tended to acknowledge the importance of Mallon's work, but still remain bound to the schema imposed by the names of scripts.

The result of this historical development, then, is that palaeography is a discipline with inadequate nomenclature that has changed little since it was canonized by the *Nouveau traité*. This nomenclature does not reflect—indeed is often a barrier to understanding—how scripts are related to each other and how they evolved. But, as the refusal to confront the challenges raised by Mallon shows, this nomenclature seems to be the tool with which palaeographers choose to work, even though problems have been recognized for almost a century, as we see from Lowe's problems with his list of half-uncial manuscripts. Nomenclature is especially problematic in the greatest period of diversity and change, i.e. in the period between Roman and Carolingian scripts. It would be easy enough to agree with Brown that the names are silly and we all know what we mean anyway, but perhaps a more fruitful way forward is to recognize that the challenge of deciding what to observe and measure is a moment of opportunity for palaeographers.

4 A Possible Course for the Future

The future for the discipline that Bernhard Bischoff foretold, if taken to its extreme, seemed to imply palaeography without palaeographers: the work would be done by measuring machines ingesting manuscripts at one end and producing analyses and transcriptions without human intervention at the other. As I have argued above, such a machine would be impossible to build because we could not always tell it what to observe. As we have also seen, however, palaeography is not a monolithic discipline; instead, there are two fundamentally different views of what palaeographers ought to be doing. On the one hand, Bischoff represents a Linnaean approach that lays great emphasis on classifying scripts and recognizing developments within a single type of script. This view of palaeography goes back to Mabillon and sees palaeography as a *Hilfswissenschaft*—its goal is to localize manuscripts in time and place and to transcribe their texts correctly so that these results can be used in larger historical narratives. Mallon, on the other hand, represents a Darwinian approach, which sees the fundamental questions as explaining the evolution of scripts and their relationships to each other, as well as looking for the mechanisms that explain these phenomena. In this view, palaeography is less concerned to deliver results to other disciplines than it is to discover its own rules to explain the ways that scribes created and modified scripts. While the Linnaeans use their classification schemes to locate the creation of a manuscript at a single point in time and space, the Darwinians are more interested in placing a given script within the internal logic of script evolution and less concerned with mapping it onto a time and place in the past.

In order to talk about the future of palaeography in general, as well as the ways that computers can be used in it, it is important to distinguish these two approaches, even though they are by no means mutually exclusive. No matter how one views palaeography and its aims, technology has certainly offered something to everyone. As noted above, the most significant contribution to the study of manuscripts of the digital age has been the enormously expanded access to an enormously expanded number of manuscripts. In addition, the use of such techniques as multispectral imaging have made it possible to read writing that was previously illegible. If nothing else, then, technology has enlarged the corpus of writing considered by palaeographers in just the same way that etching and photography had earlier. In addition, the analysis of these images holds out the possibility that many tasks of the palaeographer could be automated.

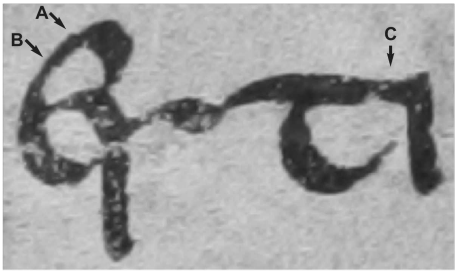


Figure 3. These four letters begin the word *estimo* in the preface of Adomnán of Iona's *Life of St Columba.* Although computer analysis often focuses on the static shapes of letters, note how much information would be lost in such an analysis. We can tell, for example, that the stroke that forms the bow of the **e** begins at Arrow A, while the stroke forming the loop begins at Arrow B. Note, too, that at Arrow C, the **t** and **i** are written as separate letters, not as a ligature. (Schaffhausen, Stadtbibliothek, Gen. 1, p. 1a.14).

In fact, it is more in the analysis of these images that the two approaches to palaeography require different approaches from computers. For the Linnaeans, automated techniques offer the ability to compare hands writing scripts of the same type. For example, given a corpus of manuscripts in Carolingian minuscule, machine analysis could perhaps be able to identify the hands of individual scribes or perhaps to create subcategories within the corpus based on the frequency of variant letterforms—just the sort of project that Derolez praised and just the sort of project that seems to represent the mainstream of automated approaches. Indeed, if this approach were applied to the range of manuscripts that Lowe considered for his list of half-uncial manuscripts, perhaps it would be possible to develop criteria that grow out of the corpus rather than being imposed upon it. Yet it is the Darwinians who may have the most interesting collaboration with technology. One fruitful approach for digital analysis, for example, could be attempting to analyze letterforms into their component strokes and pen angles. To take an extremely elementary example, Fig. 3 shows a group of eight strokes composing four letters (**esti**) from the first page of the early-eighth century manuscript of Adomnán of Iona's *Vita Columbae* now in the Stadtbibliothek of Schaffhausen in Switzerland. When we look closely, we see that these graphic signs are related in interesting ways: the strokes that make up the **e** and **s** have been modified and combined to form a ligature and the slight irregularities in the **e** marked by the Arrows A and B represent the beginning of the left-hand loop and the bow respectively of the **e**; the horizontal stroke of the **t** begins at the same place that the final stroke of the **s** ends, while the **t** and **i** might appear to form a ligature, but the way the strokes join at Arrow C show that the strokes of the two letters have not been modified, they are simply tightly spaced. No human palaeographer has the patience to go through an entire manuscript, much less an entire body of manuscripts, at such a level of detail, but this detailed description shows the amount of information to be gained from the close analysis of a fairly typical set of letters in a minuscule script—without even mentioning pen angle and line width. In other words, manuscripts have the potential to deliver up a vast quantity of information about how scribes wrote. Perhaps the automated analysis of script will soon turn its attention to reconstructing the motion of the scribe's pen on the page and following Mallon's lead explore the ways that these strokes evolved. It is then that we will begin to be able to measure the scribe's art.

Bibliography

Bischoff, Bernhard. *Paläographie des Römischen Altertums und des abendländischen Mittelalters*. 3. Aufl. Grundlagen der Germanistik 24. Berlin: E. Schmidt, 2004.

Bischoff, Bernhard. *Latin Palaeography: Antiquity and the Middle Ages*. Trans. Dáibhí Ó Cróinín and David Ganz. Cambridge: Cambridge University Press, 1990.

Brande, William Thomas and George W. Cox. *A Dictionary of Science, Literature, and Art*. London: Longmans, Green, 1867.

Brown, Julian. "Names of Scripts: A Plea to All Medievalists. Opening Address to the Oxford International Symposium on the 'Role of the Book in Medieval Culture'." *A Palaeographer's View*. Eds. Janet Bately, Michelle P. Brown, and Jane Roberts. London: Harvey Miller, 1993. 39–45.

Derolez, Albert. *The Palaeography of Gothic Manuscript Books*. Cambridge studies in palaeography and codicology 9. Cambridge: Cambridge University Press, 2003.

Foerster, Hans Philipp and Thomas Frenz. *Abriss der Lateinischen Paläographie*. Bibliothek des Buchwesens 15, 3. Überarb. und um ein Zusatzkapitel erw. Aufl. Stuttgart: Anton Hiersemann, 2004.

Gasparri, Françoise. *Introduction à l'histoire de l'écriture.* Ouvrages de référence pour l'étude de la civilisation médiévale. [Turnhout]: Brepols, 1994.

Koss, Juliet. "On the Limits of Empathy." *Art Bulletin–New York* 88.1 (2006): 139–57.

Lowe, Elias Avery. "A hand-list of half-uncial manuscripts." *Miscellanea Francesco Ehrle.* Vol. 4. Vatican City: Bibliotheca Apostolica Vaticana, 1924.

Mabillon, Jean. *De re diplomatica libri VI - Ed. 2 ab ipso auctore recognita, emendata et aucta.* Paris 1709. SBB-PK Signatur: Bibl. Diez fol. 615. <http://141.20.85.26/mabillon/>

Mallon, Jean. "Le problème de l'évolution de la lettre." *Arts et Métiers graphiques* 59 (1937): 25–30.

Mallon, Jean. *Paléographie romaine.* Madrid: Consejo Superior de Investigaciones Científicas, Instituto Antonio de Nebrija, de Filología, 1952.

Mallon, Jean. *De l'écriture. Recueil d'études publiées de 1937 à 1981.* Paris: Editions du Centre national de la recherche scientifique, 1982.

Mallon, Jean, Robert Marichal Robert, and Charles Perrat. *L'Ecriture latine de la capitale romaine à la minuscle.* Paris: Arts et métiers graph, 1939.

Marichal, Robert. "De la capitale romaine à la minuscule." *Somme typographique.* Ed. Marius Audin. 2 vols. Paris: Dupont, 1948. 1:63-111.

Martin, Henri-Jean and Vezin, Jean. *Mise en page et mise en texte du livre manuscrit.* Paris: Éditions du Cercle de la librairie – Promodis, 1990.

Oxford English Dictionary <http://dictionary.oed.com/cgi/entry/50074152?single=1&query_type=word&queryword=empathy&first=1&max_to_show=10>.

Schaffhausen, Stadtbibliothek, Gen. 1. <http://www.e-codices.unifr.ch/en/sbs/0001>.

Traube, Ludwig. *Zur Paläographie und Handschriftenkunde.* Munich: Beck, 1909.

Toustain, Charles François, and René Prosper Tassin. *Nouveau traité de diplomatique.* Vol. 2. Paris: Guillaume Desprez, 1755.

«Graphoskop», uno strumento informatico per l'analisi paleografica quantitativa

Maria Gurrado

Riassunto

«Graphoskop» è uno strumento paleografico elaborato al fine di agevolare il lavoro critico dello storico della scrittura. Il plug-in, concepito come un'estensione del software open source «*ImageJ*» è in grado di rilevare dati quantitativi di tipo metrologico a partire da una rappresentazione digitale di una scrittura data. Il metodo quantitativo è quindi applicato per la valutazione complessiva di un singolo manoscritto (analisi paleografica e analisi della *mise en page*) e,allo stesso tempo, per la valutazione di un *corpus* più esteso. Il Graphoskop, infatti, esegue alcuni calcoli statistici di base sui dati raccolti e registra automaticamente tali informazioni su un foglio di calcolo. La messa a punto della metodologia quantitativa rende possibile l'utilizzo di osservazioni analitiche per studi di carattere sintetico. La procedura è quindi orientata alla ricostruzione di un panorama più ampio: identificare la nascita di un dato fenomeno e disegnarne la curva della diffusione, per potersi interrogare più chiaramente sulle cause e modalità di tale evoluzione.

Zusammenfassung

»Graphoskop« ist ein Software-Tool zur wissenschaftlichen Untersuchung von historischen Schreibhänden. Als Plug-in für die OpenSource-Software »*ImageJ*« ist es in der Lage, spezifische Messungen an digitalen Faksimiles von Handschriften vorzunehmen und statistische Daten zu erheben. Dabei können Schrift und Schriftspiegel sowohl einzelner Handschriften als auch ganzer Handschriften-Korpora ausgewertet werden. Die Darstellung der statistischen Datenauswertung erfolgt tabellarisch. Die Anwendung statistischer Methoden auf der Basis von quantifizierbaren Daten ermöglicht analytische Beobachtungen aus einer synthetischen Perspektive. Sie ist zunächst auf allgemeine Aussagen über das erstmalige Auftreten eines bestimmten paläographischen Phänomens und dessen Verbreitung ausgerichtet. Daran anschließend kann nach den Gründen und Umständen einer solchen Entwicklung gefragt werden.

Abstract

"Graphoskop" is a software tool for the critical examination of historical handwriting. Developed as a plug-in for the open-source software "*ImageJ*", it may be used for gathering quantitative data by taking specific measurements from digital facsimiles

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 251–259.

of handwritten material. It makes it easier and quicker to apply quantitative methods—concerning scripts as well as the layout of whole pages—to descriptions of single items or surveys of larger quantities, by subjecting the collected data to basic statistical operations and exporting the results to a spreadsheet. Such quantitative methods make it possible to use analytical observations in a synthetic perspective, so as to assess general developments, e.g. by tracing the emergence and subsequent dissemination of any given graphic phenomenon, before asking questions as to why and how it occurred.

1 Introduzione

Il «Graphoskop» è uno strumento paleografico, concepito come un *plug-in* del *software open source* «*ImageJ*» (Rasband).[1] La sua funzionalità principale è quella del rilievo di dati quantitativi di tipo metrologico a partire da una rappresentazione digitale: distanza tra due linee, area di una data superficie, ampiezza di angoli diversi. Ideato come un supporto per il paleografo, può essere utilizzato sulla maggior parte dei formati immagine: TIFF, GIF, JPEG, BMP, DICOM, FITS e «raw» (8-*bit*, 16-*bit* e 32-*bit*).

La versione attuale «1.0» è di tipo beta: diverse importanti modifiche sono previste per la prossima versione. In particolare si prevede di realizzare un'interfaccia interamente personalizzabile dall'utente secondo i bisogni propri alla sua ricerca.

ImageJ è un *software* libero, originariamente sviluppato per applicazioni biomediche: è basato sul linguaggio Java ed è perciò multipiattaforma. È eseguibile sia come *applet on line* che come applicazione scaricabile su qualsiasi tipo di sistema operativo (Linux, Mac OS X, Windows, etc.): più di 400 *plugins* sono attualmente disponibili gratuitamente *on line*.

La scelta di un *software open source* ha ragioni multiple: dalla gratuità della licenza all'idea, seducente, di collaborare allo sviluppo di un sistema pur lavorando da soli. La ragione principale resta in ogni modo la flessibilità, l'architettura libera del *software* che garantisce la possibilità di personalizzare le applicazioni e di essere responsabili dell'evoluzione del proprio strumento di lavoro. Il Graphoskop, infatti, è uno strumento di lavoro, uno dei tanti che a volte si immaginano soltanto e che non si ha mai il tempo di portare a termine. L'idea è semplice e nasce da un bisogno concreto della ricerca: in questo caso, dell'*expertise* paleografica.

Chiunque si sia cimentato nell'analisi paleografica sa che l'esame di una scrittura richiede una certa elasticità mentale: le linee del testo sono raramente perfettamente

[1] Definire la paternità del Graphoskop non è cosa semplice. Se è vero che l'idea del *plug-in* appartiene a chi scrive, una buona parte delle funzionalità da esso previste sono il frutto di lunghe e interessanti conversazioni con Marc Smith e Denis Muzerelle. Il Graphoskop, infine, è stato interamente sviluppato da Giancarlo Lestingi per l'École Nationale des Chartes.

parallele, l'estensione delle aste ascendenti e discendenti non è mai perfettamente regolare, il modulo delle lettere non è mai identico e cosi via. La frequentazione dei manoscritti ci ha abituati a considerare costantemente un margine di imprecisione nella valutazione generale sia della tipologia della scrittura che del livello d'esecuzione della stessa. Benché spesso non precisabile, la cosiddetta «impressione generale» ha molto spesso un ruolo essenziale nell'esame di una scrittura. Questo margine d'imprecisione, non sempre quantificabile o esplicitabile, è probabilmente l'ostacolo più importante all'analisi automatica e informatizzata della scrittura e, più generalmente, costituisce forse il problema più delicato della stessa analisi paleografica (Gilissen). Armando Petrucci (p. X) ha sottolineato come «in generale, le misurazioni e complessivamente l'approccio quantitativo non sembrano costituire una tecnica adatta allo studio ed alla conoscenza di fenomeni culturali di tale mutevolezza e sottigliezza quali la scrittura a mano.»

2 Funzionalità

Graphoskop si vuole uno strumento su misura, un ausilio nel rilievo di distanze fra punti accuratamente scelti dall'utente e indicati manualmente e singolarmente su ogni immagine: distanza tra il principio e la fine di un'asta, tra una parola e l'altra, altezza del corpo della scrittura, angolo di scrittura, inclinazione delle aste e quant'altro si voglia misurare. Si tratta quindi di uno strumento prevalentemente descrittivo che coniuga le scelte dell'utente all'automatizzazione dei calcoli ed alla registrazione dei risultati.

Il *plug-in*, del resto, è nato nel contesto di una tesi dell'École nationale des chartes finalizzata allo studio delle scritture corsive librarie nel XIII e XIV secolo. Non è il frutto di un progetto informatico di gran respiro, né ha l'ambizione di rilevare indizi utili alla datazione o localizzazione delle scritture. L'adozione di un protocollo solo parzialmente automatizzato delle procedure di analisi grafica costituisce un complemento allo sguardo critico dell'utente e, allo stesso tempo, una garanzia per lo storico, solo responsabile dell'analisi paleografica.

Graphoskop esegue dei calcoli statistici di base sui dati raccolti: media, moda e deviazione standard si rivelano informazioni utilissime non solo per la valutazione complessiva di un singolo manoscritto ma anche, e soprattutto, per la valutazione di un *corpus* più esteso. La rappresentatività e l'attendibilità del campione analizzato incombono esclusivamente allo storico della scrittura.

Un altro tipo di misura rilevato dal *plug-in* è la percentuale di *pixels* bianchi e neri in una superficie delimitata dall'utente (*Region Of Interest* - ROI). A questo scopo Graphoskop crea una copia della ROI in memoria, la binarizza e calcola la percentuale dei *pixels*. In questo modo possono essere facilmente misurati sia il «coefficiente di riempimento» della pagina che il «coefficiente di sfruttamento» della stessa (Agati).

Nel secondo caso, ad esempio, se l'utente ha avuto cura di selezionare esclusivamente alcuni righi di scrittura, escludendo iniziali o elementi decorativi, tale risultato si traduce con la percentuale di scrittura (*pixels* neri) sul supporto (*pixels* bianchi). Lo stesso tipo di calcolo può essere effettuato sul foglio intero («coefficiente di riempimento») o sul corpo della scrittura: in quest'ultimo caso, peraltro, si può rivelare utile calcolare la media di più selezioni di righi di scrittura sulla stessa pagina.

A questo proposito pare opportuno ricordare come già Léon Gilissen avesse inteso l'importanza del «peso» della scrittura nel contesto dell'*expertise* paleografica. Nel capitolo ad esso consacrato, egli aveva descritto un calcolo laborioso, nonché mai apertamente contestato, per la misura del «poids de l'écriture». Le sue formule avevano l'interessante vantaggio di relativizzare ogni elemento concorrente all'aspetto generale della scrittura. Gilissen stesso, tuttavia, aveva indicato il carattere non assoluto di tali formule, in special modo riguardo alle scritture «filiformi» e a «*toutes les écritures dont la progression est obtenue artificiellement et uniquement par des traits obliques maigres*» (35). Marc Smith aveva del resto avuto modo di suggerire la sostituzione dei complessi calcoli di Gilissen con il rilievo automatico in *pixels* della densità relativa della scrittura (Smith nota 15). Graphoskop funziona dunque alla stregua di una scorciatoia attraverso tali inconsueti calcoli: una semplificazione che produce risultati non identici ma connessi e, a prima vista, non meno pertinenti.[2]

Il tipo di registrazione dati proposta dal *plug-in* si giustifica anch'esso con le esigenze della ricerca, in particolar modo con la funzionalità di una base di dati. Il foglio di calcolo si rivela un modo semplice e diffuso di gestire i dati di una ricerca. Nel contesto di uno studio paleografico, il confronto tra numerosi esempi di scritture è manovra ordinaria e imprescindibile: oltre a garantire la visualizzazione immediata dei dati, il foglio di calcolo costituisce una facile scorciatoia verso la rappresentazione grafico-statistica degli stessi. Nel caso specifico di Graphoskop un'ulteriore funzionalità permette di registrare di volta in volta e sullo stesso foglio le misure rilevate sull'intero *corpus*. Il *plug-in* genererà del resto un foglio di calcolo per ogni immagine analizzata.
Riassumendo, il *plug-in*:

- calcolerà la media delle distanze (riguardanti ad esempio il corpo della scrittura come indicato per terreno d'indagine da Muzerelle (34)) che l'utente avrà tracciato su una singola immagine;
- misurerà la dispersione dei valori attorno a quest'ultima permettendo di valutarne rappresentatività e affidabilità (deviazione standard);
- isolerà il valore più ricorrente;
- calcolerà la densità di pixels bianchi e neri (i.e. la densità della scrittura);

[2] Fra i primi esperimenti effettuati sulle scritture medievali e comprendenti il computo dei pixels bianchi e neri cf. Friedman 1994.

- registrerà tutte le misure rilevate ed i dati statistici ottenuti su diversi fogli di calcolo.

3 Interfaccia

L'interfaccia grafica di Graphoskop è composta dalla barra di applicazioni di *ImageJ*, di una finestra «*results*» secondaria e della finestra propria al *plug-in*:

- la finestra principale, contenente l'immagine da analizzare, costituisce il punto di controllo di tutto il *plug-in*. Essa è corredata da due righelli, uno orizzontale ed uno verticale, settati secondo un'unità di misura a scelta dell'utente (*pixel*, mm o cm).

Per difetto, Graphoskop considera che le immagini da analizzare siano in scala reale; tuttavia è possibile settare la scala a scelta dell'utente: basta tracciare una linea direttamente sull'immagine ed attribuirle un valore numerico perchè i righelli si adattino alla nuova scala. In assenza della consueta riga accostata al manoscritto fotografato, questa funzionalità permette di lavorare su qualsiasi immagine di cui si conosca almeno una dimensione (per esempio la lunghezza di una parola, il diametro di un timbro, la larghezza di un'intercolonna, etc.). Avendo a disposizione una fotografia di una pagina intera, è possibile peraltro rilevare non solo dati relativi alla scrittura, ma anche alla *mise en page* del manoscritto (interlinea, margini, intercolonna, etc.):

- la finestra secondaria mostra i calcoli effettuati dal *plug-in* in tempo reale. Ciò permette di avere un riscontro immediato delle operazioni in corso e un'anteprima di tutte le misure di uno stesso tipo che saranno poi registrate nel foglio di calcolo.

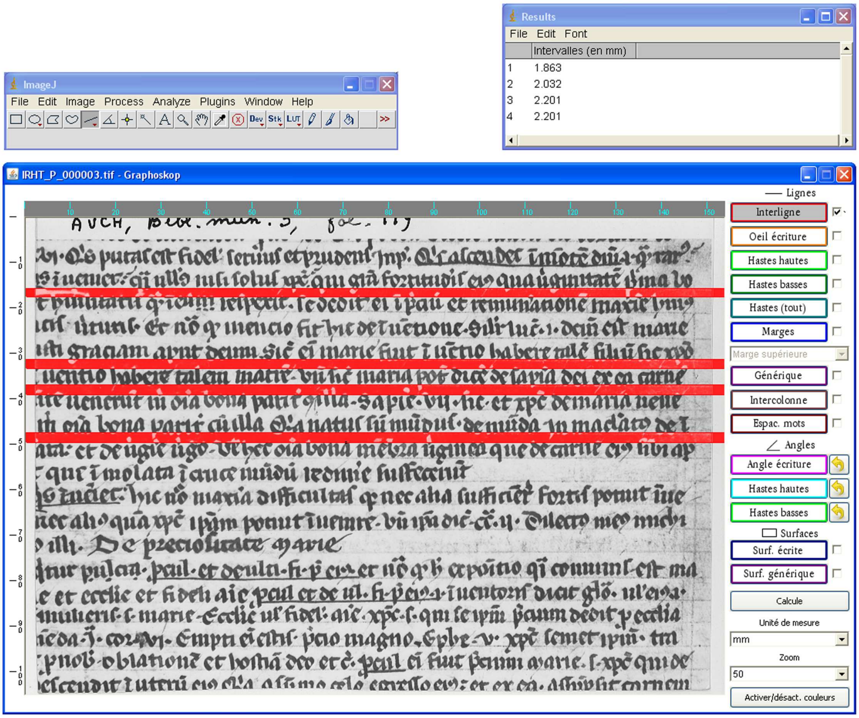
4 Tasti

Il *plug-in* dispone di tre tipi di strumenti di selezione: linea, angolo e rettangolo.

Tutti e tre sono di volta in volta, e senza limite di numero, tracciati dall'utente secondo le sue esigenze. Ogni strumento è attivato da un tasto di colore diverso che porta il nome della distanza da misurare. Ogni tasto è accompagnato da un comando che permette di attivare una funzione specifica di controllo della zona selezionata. Nell'attesa di comandi personalizzabili, la versione 1.0 presenta 17 tasti[3], ripartiti nel modo seguente:

- «linea orizzontale»: interlinea, aste ascendenti, aste discendenti, distanza aste asc./disc., corpo della scrittura, margine superiore, margine inferiore, distanza generica;

[3] La «distanza generica» costituisce un unico tasto declinabile in «linea orizzontale» o «linea verticale» (per qualsivoglia zona non specificata).



	Intervalles (en mm)
1	1.863
2	2.032
3	2.201
4	2.201

Unité de mesure: mm

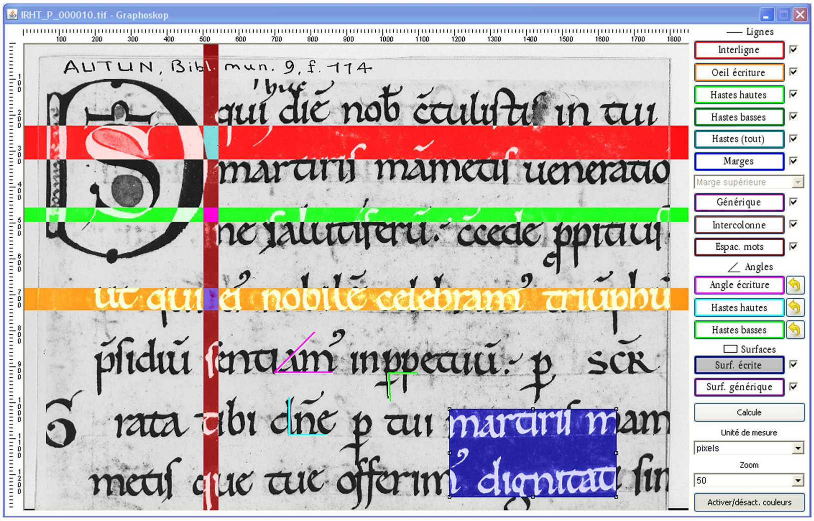
Zoom: 50

Figura 1.

- «linea verticale»: spazio fra parole, margine interno, margine esterno, intercolonna, distanza generica;
- «angolo»: inclinazione aste ascendenti, inclinazione aste discendenti, angolo di scrittura (ossia inclinazione della penna);
- «rettangolo»: specchio di scrittura, superficie generica.

L'utente posiziona dunque le linee (*drag and drop*), traccia gli angoli e delimita una o più ROI; infine clicca sul tasto «calcola».

Una nuova finestra permette la richiesta di calcoli più complessi: media, deviazione standard, moda, area e densità. Per difetto, Graphoskop rileva tutte le misure di distanze fra linee e calcola l'ampiezza degli angoli tracciati sull'immagine.



Unité de mesure: pixels

Zoom: 50

Figura 2.

5 Registrazione dati

Al momento del primo salvataggio dei calcoli, vengono creati due fogli di calcolo: un foglio complessivo ed uno proprio ad ogni immagine.

1. Il foglio «complessivo»:

Nella prima colonna sono registrati i nomi di ogni *file*. Se l'immagine è stata precedentemente chiamata con il riferimento preciso al manoscritto fotografato, in questa colonna saranno registrati via via tutte le collocazioni dei manoscritti costituenti il *corpus* da analizzare.

Le altre colonne portano il nome dei tasti presenti nella finestra principale. Più precisamente il *plug-in* calcola:

- media, deviazione standard e moda di: interlinea, aste ascendenti, aste discendenti, distanza aste ascendenti/discendenti, spazio fra parole, angolo di scrittura, inclinazione aste ascendenti, inclinazione aste discendenti e altezza del corpo della scrittura;
- media dell'intercolonna;
- densità (*pixels* bianchi/neri) di: specchio di scrittura, superficie generica e corpo della scrittura;
- base, altezza ed area del rettangolo di: superficie generica e specchio di scrittura.

2. Il foglio specifico dell'immagine analizzata:

Graphoskop crea tanti fogli distinti quante sono le immagini analizzate. Ognuno di questi fogli porta il nome del file trattato. Su ciascuno di questi fogli saranno registrate tutte le misure che l'utente avrà richiesto e che il *plug-in* avrà mostrato di volta in volta nella finestra secondaria. In guisa di sintesi, nelle ultime righe di ogni foglio l'utente ritroverà i dati statistici (media, deviazione standard e moda) presenti nel foglio complessivo.

È inoltre possibile salvare le foto trattate, conservando in evidenza le selezioni effettuate, in una cartella scelta dell'utente. Tale procedura permetterà una verifica puntuale non solo del metodo scelto, ma della sua applicazione.

L'impiego di programmi semplici per il rilievo metrologico delle scritture medievali potrebbe apportare oggi un risparmio di tempo considerevole per uno studio comparato di varie centinaia di esemplari, rendendo così possibile la messa alla prova di metodi descrittivi analoghi a quelli prospettati nel 1973 da Léon Gilissen, ma rimasti da allora lettera morta o quasi, principalmente per la scarsa fattibilità pratica.

Si tratta, in sostanza, di un affinamento della tecnica di expertise che permette, oggi, di travalicare l'analisi di un campione ristretto e di sondare le possibilità di rilievi sistematici su un gran numero di esemplari. I risultati forniti da Graphoskop intendono agevolare il lavoro critico del paleografo, del codicologo o anche del diplomatista supportando sia l'analisi paleografica in senso stretto che l'analisi della *mise en page*. L'analisi dell'interazione tra testo, scrittura e supporto permetterà di far luce sulle soluzioni adottate dai copisti per garantire allo stesso tempo leggibilità ed economia dello spazio sulla pagina. Si potrebbero ad esempio valutare le variazioni *inter* e *intra specimina*:

- dell'interlinea in funzione delle proporzioni delle colonne;
- delle proporzioni delle aste e del corpo della scrittura in relazione al modulo di quest'ultima (per esempio nel passaggio dalla scrittura carolina alla *littera textualis*);
- dell'estensione delle aste discendenti e/o ascendenti (in alcune scritture caroline: «p» più estesa di «q», estensione di «f» assimilabile a «r» discendente sotto il rigo di base);
- dell'angolo di scrittura e/o dell'inclinazione delle aste (nelle scritture caroline esaminate, l'inclinazione delle aste discendenti – «p» e «q» – è più costante di quella delle aste ascendenti – «l» e «d»).

Com'è noto, la messa a punto della metodologia quantitativa rende possibile l'utilizzo di dati analitici per studi di carattere sintetico. Si tratta quindi di superare lo stadio di pochi esemplari giudicati rappresentativi in determinati contesti per cercare di ricostruire un paesaggio più vasto, senza alcuna pretesa di esaustività: identificare la nascita di un dato fenomeno, disegnarne la curva della diffusione ma anche, e soprattutto, interrogarsi sulle cause e modalità di tale evoluzione.

Graphoskop è stato concepito come supporto all'*expertise* paleografica: non un «occhio del paleografo» informatizzato, bensì un modo di automatizzare la parte tecnica dell'analisi. In nessun caso si vuole dunque incoraggiare l'adozione di un protocollo che permetta l'automatizzazione dell'analisi storica senza rimettere in causa i principi sui quali essa stessa riposa.

Bibliografia

Agati, Maria Luisa. *Il libro manoscritto, introduzione alla codicologia.* Roma: L'Erma di Bretschneider, 2003. 237–238.

Friedman, John B. Some contour features in medieval script: a preliminary study. *Advances in handwriting and drawing. A multidisciplinary approach.* Eds. C. Faure, P. Keuss, G. Lorette, and A. Vinter. Parigi: Europia, 1994. 547–560.

Gilissen, Léon. *L'expertise des écritures médiévales: recherche d'une méthode, avec application à un manuscrit du XIe siècle, le lectionnaire de Lobbes, codex Bruxellensis 18018.* Gand: Editions scientifiques E. Stiry-Scienta, 1973.

Muzerelle, Denis. Le geste et son ombre: essai sur le «rapport modulaire» des écritures. *Gazette du livre médiéval* 35 (1999): 32–45.

Petrucci, Armando. Preface. *La face cachée du livre médiéval. L'histoire du livre vue par Ezio Ornato, ses amis et ses collègues.* Roma: Viella, 1997. I–XXVI.

Rasband, Wayne. *ImageJ. Image Processing and Analysis in Java.* Research Services Branch, National Institute of Health. 2006ff.
<http://rsb.info.nih.gov/ij/index.html>.

Smith, Marc. Numérisation et paléographie. *Le médieviste et l'ordinateur* 40 (2001).
<http://lemo.irht.cnrs.fr/40/mo40-03.htm>.

Forschung am Rande des paläographischen Zweifels: Die EDV-basierte Erfassung individueller Schriftzüge im Projekt DA*mal*S[*]

Wernfried Hofmeister, Andrea Hofmeister-Winter, Georg Thallinger

Zusammenfassung

Das Pilotprojekt DA*mal*S (Datenbank zur Authentifizierung mittelalterlicher Schreiberhände) hat es sich zum Ziel gesetzt, neue Kriterien zur Unterscheidung von Schreiberhänden in mittelalterlichen deutschsprachigen Handschriften aufzustellen und zuverlässigere Methoden und Werkzeuge für diese Aufgabe zu entwickeln. DA*mal*S beruht auf den drei Säulen einer elementgetreuen Basistransliteration in XML, computerbasierten graphetischen Analysen und einem neuartigen Verfahren der bildorientierten Mustererkennung. Diese Säulen sind in eine Datenbankstruktur integriert, welche sowohl die Archivierung als auch die technisch hochkomplexe Verarbeitung der Bild- und Textdokumente leistet. Auf diese Weise bietet DA*mal*S eine Art Brille, durch die paläographische ExpertInnen bei ihrer Schriftbegutachtung unterstützt werden. In einer weiteren Ausbaustufe soll das Projekt DA*mal*S in ein neues Projekt namens MOSES (Musterorientiertes System zur Erfassung von Schriftindividualität) eingebettet werden, welches sich auf neuzeitliche und aktuelle handgeschriebene Materialien ausdehnen lässt, um dann z. B. auch für forensische Zwecke hilfreich zu sein.

Abstract

In order to provide objective criteria for distinctions between presumably different writing hands in medieval German vernacular manuscripts the project DA*mal*S (Datenbank zur Authentifizierung mittelalterlicher Schreiberhände/Database for the Authentication of Medieval Writing Hands) has developed new methods and tools: three pillars—a palaeographically extremely detailed XML transliteration, manifold graphetical statistics and image-based pattern recognition—have been integrated into an innovative database as complex and highly interrelated techniques for the analysis of handwritten documents. By these means DA*mal*S also offers "virtual spectacles"

[*] DA*mal*S steht für ›Datenbank zur Authentifizierung mittelalterlicher Schreiberhände‹. Die nachfolgenden Ausführungen lassen sich wie folgt grob den drei Beiträgern zuordnen: Wernfried Hofmeister: Abschnitt 1, 4.1 u. 5; Andrea Hofmeister-Winter: 2.1 u. 3.1; Georg Thallinger: 2.2, 3.2 u. 4.2.

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 261–292.

through which palaeographic experts may look and thereby be supported in their challenging judgements. In a further step DA*mal*S is to be incorporated into a new project called MOSES (Musterorientiertes System zur Erfassung von Schriftindividualität/Pattern Orientated System for the Detection of Individuality in Handwriting) and e. g. also be helpful in solving current forensic problems.

1 Vorbemerkungen – DA*mal*S im Kontext einer neuen Überlieferungsphilologie

Die Frage nach der Anzahl der Hände, die an einem Überlieferungsträger gearbeitet haben, gehört für die Paläographie seit jeher zu den wichtigsten, dabei zugleich zu den herausforderndsten, und sie hat in jüngster Zeit noch an Brisanz gewonnen: Gestärkt durch die Bestrebungen der ›New Philology‹, bemüht man sich in zahlreichen historischen Textfächern um eine umfassende Würdigung jedes einzelnen Aufzeichnungsprozesses rund um die vielfältigen Aspekte von ›Produktion und Kontext‹.[1] Motiviert wird diese erhöhte Aufmerksamkeit für die individuelle Genese und für die nur scheinbar banale Text-Materialität[2] aller Schriftdokumente durch ein neues *Werk*verständnis, welches sich nicht zuletzt darin ausdrückt, dass man umfassender als zuvor bereit ist, eine prinzipielle Werk-›Offenheit‹ anzuerkennen: Anstatt weiterhin eine imaginäre ›Urfassung‹ in den Mittelpunkt zu stellen, welche es anhand aller präsumtiv minderwertigen Überlieferungsträger zu rekonstruieren gelte, hat man erkannt, dass viele der erhaltenen Überlieferungen als ernst zu nehmende *Varianten* zu sehen sind, oft als ganz eigenwertige *Fassungen*, welche an veränderte Rezeptionsbedingungen angepasst worden waren[3] und eine eingehende Neubewertung verdienen. Alle Werk-Handschriften sind im Grunde für uns damit von philologischem Rohmaterial zu aussagekräftigen Dokumenten geworden. In besonderem Maße gilt dies dort, wo bei näherer Betrachtung schon in historischer Zeit eine systematische Überlieferungs*strategie* sichtbar wird, mithin ein fast präphilologisch zu nennendes Konzept, welches sich in dem Maße ausdrückt, in welchem von den (ab-)schreibenden Persönlichkeiten unter Bedachtnahme auf diverse

[1] Unter diesem Titel fand 1998 in Den Haag die 7. internationale Tagung der Arbeitsgemeinschaft für germanistische Edition statt: Vgl. Produktion und Kontext.

[2] »Materialität in der Editionswissenschaft« lautete das Generalthema der 12. internationalen Tagung der Arbeitsgemeinschaft für germanistische Edition vom 13.–16. Februar 2008 an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) in Berlin. (Näheres findet sich dazu auf der Website der AG.)

[3] Besonders schöne Beispiele für die Eigenwertigkeit von Varianten und Fassungen konnten im Rahmen des ›Parzival-Projekts‹ von Michael Stolz erzielt werden. Aktuelles dazu auf der Projekt-Homepage.

Faktoren wie Verfügbarkeit, Vollständigkeit, Verlässlichkeit der Überlieferungsvorlagen resp. Quellen gearbeitet wurde, sodass im Grunde frühe *Editionen* entstanden.[4]

Es ist also diese neue Sicht auf den potenziellen Eigenwert jeder einzelnen Überlieferung, die auch unser Interesse an den dafür verantwortlichen *SchreiberInnen* hat wachsen lassen, weil ja letztlich *sie* es waren, die sich im Zuge einer solchen Text*re*produktion mit eingeschrieben haben – mit all ihren (mehr oder minder gewollten) Textabweichungen in Gestalt von Kürzungen, Wortumstellungen, Begriffsauswechslungen etc. Entsprechend bedeutsam ist es für die gesamte Textforschung zu wissen, ob bzw. wo wir es in einem Codex mit ein und derselben Hand (im Sinn von ein und derselben Person) zu tun haben, weil diese ja gemäß mittelalterlicher ›Überlieferungsmoral‹ eine spezifische *inhaltliche* und *sprachlich-formale* Textverantwortung[5] trägt und diese auch durch ihre Arbeit zum Ausdruck bringt: Neben den schon angedeuteten Strategien im *Inhalts*bereich (in Bezug auf die rein ›stoffliche‹ Wiedergabe von schriftlichen Vorlagen oder mündlichen Diktaten) sind es die *linguistisch* relevanten Aspekte, welche sich besonders offensichtlich an ein schreiberspezifisches Sprachvermögen rückgebunden zeigen und uns daher nur unter Bedachtnahme auf diesen Zusammenhang als Evidenz für sprachhistorische Erscheinungen dienen können. Oder anders ausgedrückt: Gerade in der mittelalterlichen Zeit, wo das schriftliche Überliefern von Texten bekanntlich eine Fertigkeit darstellte, die in hohem Maße von frei migrierenden und deshalb an verschiedenen Orten arbeitenden SpezialistInnen erledigt wurde, ist es für unsere Erschließung der historischen Laut- und Formenvielfalt von eminenter Bedeutung, die Grenzen zwischen verschiedenen Schreiberpersönlichkeiten verlässlich ausmachen zu können; erst das ermöglicht eine bewusste Berücksichtigung ihrer schriftsprachlichen Idiosynkrasien und erlaubt in weiterer Konsequenz eine fundierte(re) Debatte über das Heranwachsen von regionalen und überregionalen Sprachnormen! – In diesem Sinne spiegelt eine paläographische *ad fontes*-Bewegung, für welche die individuellen Schriftzüge – wie eine ›Grammatik der Schreiberhände‹ (Hofmeister-Winter 2005) – zum wichtigsten Anker für zahlreiche weitere Erkenntnisse geworden sind, den Urgrund einer neuen, viele Fächer verbindenden *Überlieferungs-Philologie*[6] wider.

[4] Der editorische Eigenwert mittelalterlicher Überlieferungen wurde jüngst im Rahmen der mediävistischen Fachtagung »Wege zum Text« betont. Siehe dazu den Beitrag des Tagungsleiters, Wernfried Hofmeister, der darauf in seinem »Problemaufriss zum Stellenwert von Editionen und ihrer Verfügbarkeit« Bezug nahm; diese Ausführungen sind auf der Tagungs-Homepage als Podcast nachzuhören; der Sammelband zur Tagung wird voraussichtlich im Sommer 2009 im Druck erscheinen.

[5] Die Arbeitsteiligkeit zwischen Text-Autor und Überlieferern erscheint zwar im Lichte diverser Sorgen von mittelalterlichen Autoren bezüglich drohender Textentstellungen durch schlechte Textweitergaben nicht spannungsfrei gewesen zu sein, belegt aber indirekt ebenfalls genau diesen – von den SchreiberInnen offensichtlich gerne genutzten – Freiraum für ihre Textaufzeichnung. Daneben kennen wir freilich auch Signale der Zustimmung zu den unvermeidlich ausbleibenden Texteingriffen, müssen uns jedoch in einigen dieser Fälle fragen, ob nicht eine ironische Brechung einer solchen ›Lizenz zum Verändern‹ vorliegt. Dies steht etwa auch für den Autor Hugo von Montfort zu vermuten; vgl. Hofmeister 2004.

[6] Für ihre Etablierung wird explizit von Wernfried Hofmeister (Hofmeister 2001) geworben. Aufgegriffen

1.1 Globalziele von DA*mal*S

Das nun vorzustellende Projekt DA*mal*S[7] versteht sich ganz bewusst als ein Teil dieser eingangs skizzierten Rückbesinnung auf das Elementare unserer Schriftkultur. Es ist angetreten, um dort, wo bislang mittels herkömmlicher paläographischer Befundungen zwischen einzelnen ForscherInnen kein Konsens über die Anzahl der beteiligten Schreiberhände zu gewinnen war, mit einer plurimethodischen Herangehensweise *relevante Indikatoren* für eine Art *Schriftindividualität* zu ermitteln und diese Spezifika hernach auf höchstem technischen Niveau durch objektive schriftanalytische Messwerte zu dokumentieren. Somit will dieses neue Instrumentarium nicht mehr, aber auch nicht weniger sein als eine Art *Brille*, durch die unser paläographisch geschultes Auge *zusätzlich* blicken kann, um ein intersubjektiv teilbares Urteil fällen zu können. Dies ist nicht nur hilfreich in jenen Fällen, wo verschiedene SchriftexpertInnen unterschiedliche Meinungen vertreten, sondern auch dort, wo unsere Augen nach wiederholtem Befunden einer fraglichen Handschrift zu durchaus schwankenden Ergebnissen gelangen.

Noch hat – um im Bild zu bleiben – diese DA*mal*S-Brille nicht ihren letzten Schliff erhalten, da einige ergänzende Optimierungen erst zu leisten sind, doch für ihren zentralen Probefall, nämlich den deutschsprachigen cpg 329[8] (um 1415, aus dem Besitz Hugos von Montfort), hat sie offensichtlich schon genügend Trennschärfe gezeigt, um uns dabei behilflich zu sein, eine langjährige Streitfrage erfolgreich zu lösen; ohne dazu (im Vorgriff auf die nächsten Beitragsabschnitte) die Details vorwegzunehmen, seien zunächst ganz allgemein und überblickshaft das operationale Design dieses neuen Befundungsinstruments kurz vorgestellt sowie einige seiner methodischen Hintergründe.

1.2 Arbeitsprozess-Schema für DA*mal*S: Authentifizierungs-Szenario

Als Orientierungshilfe für die folgenden Ausführungen, welche dem zentralen ›Authentifizierungs-Szenario‹ gelten, dient das ›Arbeitsprozess-Schema‹ (Abb. 1)[9]. Dessen 3-säulige Grundstruktur ist erstmals Anfang 2008 einem internationalen Fach-

und durch neue Beobachtungen an der Materialität ergänzt wurde dieser Beitrag jüngst von Martin Schubert (Schubert).

[7] Aktuelle Informationen über dieses Projekt bietet die Projekt-Homepage. – Dem Begründer von DA*mal*S, Wernfried Hofmeister, wurde zur Förderung dieser Pilotstudie von der Forschungsabteilung des Landes Steiermark im Jahr 2006 ein namhafter Förderungsbetrag zuerkannt. Ein Großteil dieser Förderung dient seinem Projektpartner, der Forschungsgesellschaft Joanneum Research (vertreten durch Georg Thallinger) für die Entwicklung der technischen Projekt-Applikationen. In Summe wirken an DA*mal*S (freilich nur zeitweise bzw. auf Werkvertragsbasis) rund 10 Personen mit, darunter VertreterInnen des Zentrums für Informationsmodellierung in den Geisteswissenschaften (ZIMig) an der Karl-Franzens-Universität Graz.

[8] Siehe die Online-Präsentation dieses Codex sowie zu seiner – für DA*mal*S grundlegenden – Einbindung in die neue Hugo von Montfort-Ausgabe von Wernfried Hofmeister die Editions-Homepage.

[9] Alle im Abbildungsteil gezeigten Ausschnitte aus dem Montfort-Codex cpg 329 stammen von der Online-Präsentation cpg 329 der UB Heidelberg.

plenum bekannt gemacht worden (Hofmeister und Hofmeister-Winter) und hat dabei über die Fachgrenzen hinweg Zustimmung erhalten. Für diesen Sammelband erfährt das soeben erwähnte ›Arbeitsprozess-Schema‹ nun außer einer differenzierteren graphischen Aufbereitung vor allem eine detailliertere Beschreibung seiner *technischen* Komponenten.[10]

Generell ist anzumerken, dass DA*mal*S von einer zweifachen Datenbankstruktur getragen wird, einer äußeren und einer inneren: Nach *außen* hin formen die einzelnen Überlieferungsträger sowie die darin ermittelten Schreiberhände Datensätze, durch deren Verwaltung (u. a.) das spätere Wiedererkennen von bereits erfassten Daten/Personen ermöglicht wird. Aber auch in ihrem *Inneren* stützt sich jede einzelne Dokument-Aufnahme und Befundung auf ein (interrelationales) Datenbank-Design, durch das alle Einträge miteinander verknüpft werden können, um ein sukzessives Sammeln und abschließendes Auslesen aller Informationen, Werte und Daten zu erlauben. Nähere Hinweise zur dafür gewählten Spezialdatenbank *imdas pro*, welche sowohl die äußere als auch die innere DA*mal*S-Struktur trägt, finden sich weiter unten in Abschnitt 2.2.

Der gesamte Authentifizierungsablauf ruht im Wesentlichen auf den schon erwähnten drei Säulen resp. Untersuchungsansätzen: Sie bestehen aus der ›klassischen‹ Paläographie, der ›graphetischen Statistik‹ und der ›musterorientierten Schriftbild-Erfassung‹. Diese drei Ansätze sind nur teilweise als sukzessiv zu denken, denn manches läuft parallel ab. Den *Beginn* des Arbeitsprozesses repräsentiert jedoch immer (in Säule 1) die kodikologische Beschreibung des Überlieferungsträgers und seiner *paläographischen* Charakteristika. Am Ende dieser Routine steht als Ziel eine maßgeschneiderte, dabei bis in die unscheinbarsten Teile elementgetreue Basistransliteration in XML-Codierung. Damit ist (für Säule 2) die Möglichkeit zur *statistischen* Bestimmung jener Graphvarianten, -teile und -sequenzen gegeben, die es via Datenbank auszuwerten gilt, um allfällige signifikante Schwankungen erkennbar zu machen. Praktisch simultan dazu findet (in Säule 3) eine messtechnische und algorithmische Befundung des *Schriftbildes* statt; für die hier nötige höchste Präzision sorgen die zuvor schon in die Datenbank integrierten *Bild-Digitalisate* aller Schriftstücke, indem deren Bildoberfläche mit allen Graph-Elementen der XML-Transliteration verknüpft wird. – Alle drei Befundungssäulen zusammen ergeben bereits einen ausgeprägten *plurimethodischen* Ansatz. Ergänzt wird dieser interdisziplinäre Zugriff durch die Einbeziehung kognitionspsychologischer Zusammenhänge (rund um die Verarbeitung und den Ausdruck gestalthafter Einheiten durch feinmotorische Impulse bei handschriftlicher Sprachreproduktion), aber auch der quantifizierenden Linguistik sowie allg. der Textmodellierung, für die es neben bzw. in Kooperation mit TEI neue Wege für noch feinere Abbildungen graphetischer ›Elementarteilchen‹ zu finden galt.

[10] Dies erfolgt in den Abschnitten 2.2, 3.2 und 4.2; bezüglich der namentlichen Zuordnung dieser (und aller übrigen) Beitragsabschnitte vgl. Anm. 1.

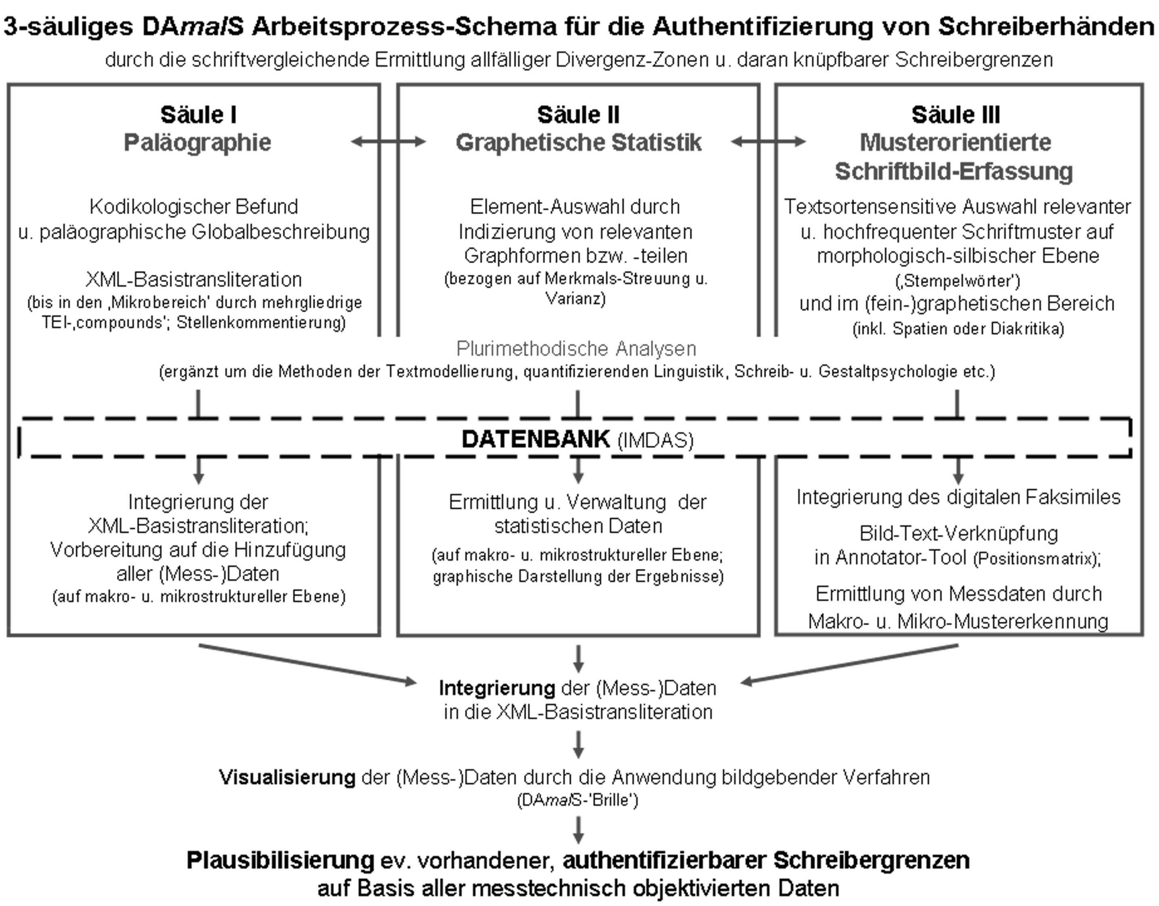

3-säuliges DA*mal*S Arbeitsprozess-Schema für die Authentifizierung von Schreiberhänden
durch die schriftvergleichende Ermittlung allfälliger Divergenz-Zonen u. daran knüpfbarer Schreibergrenzen
Säule I
Paläographie
Kodikologischer Befund u. paläographische Globalbeschreibung
XML-Basistransliteration
(bis in den ‚Mikrobereich' durch mehrgliedrige TEI-‚compounds'; Stellenkommentierung)
Säule II
Graphetische Statistik
Element-Auswahl durch Indizierung von relevanten Graphformen bzw. -teilen
(bezogen auf Merkmals-Streuung u. Varianz)
Säule III
Musterorientierte Schriftbild-Erfassung
Textsortensensitive Auswahl relevanter u. hochfrequenter Schriftmuster auf morphologisch-silbischer Ebene
(‚Stempelwörter')
und im (fein-)graphetischen Bereich
(inkl. Spatien oder Diakritika)
Plurimethodische Analysen
(ergänzt um die Methoden der Textmodellierung, quantifizierenden Linguistik, Schreib- u. Gestaltpsychologie etc.)
DATENBANK (IMDAS)
Integrierung der XML-Basistransliteration; Vorbereitung auf die Hinzufügung aller (Mess-)Daten
(auf makro- u. mikrostruktureller Ebene)
Ermittlung u. Verwaltung der statistischen Daten
(auf makro- u. mikrostruktureller Ebene; graphische Darstellung der Ergebnisse)
Integrierung des digitalen Faksimiles
Bild-Text-Verknüpfung in Annotator-Tool (Positionsmatrix);
Ermittlung von Messdaten durch Makro- u. Mikro-Mustererkennung
Integrierung der (Mess-)Daten in die XML-Basistransliteration
Visualisierung der (Mess-)Daten durch die Anwendung bildgebender Verfahren
(DA*mal*S-'Brille')
Plausibilisierung ev. vorhandener, authentifizierbarer Schreibergrenzen auf Basis aller messtechnisch objektivierten Daten

Abbildung 1. Arbeitsprozess-Schema für die Authentifizierung mittelalterlicher Schreiberhände.

Dank dieser konsequenten XML-Codierungsweise können in weiterer Folge alle hilfreichen Informationen (wie paläographisch-textkritische Stellenkommentare, abschnittsbezogene Frequenzwerte oder gestaltvergleichende Konvergenz- bzw. Divergenzparameter zwischen Graphen oder Graphfolgen/Wörtern) direkt im jeweiligen Transliterationsbereich gespeichert resp. dort hinein *integriert* werden. Danach lassen sich unter Einbeziehung der bildgebundenen Positionsdaten sämtliche Werte gleichsam in die Handschrift zurückprojizieren und – gesteuert von allen Vergleichsdaten – punktgenau jene Bildbereiche erkennen, an denen sich auffallende und evtl. nur durch einen Schreiberwechsel erklärbare Veränderungen abzeichnen: Zur Verdeutlichung der abstrakten Vergleichszahlen dienen unterschiedliche *Visualisierungstechniken*; indem sie unser Auge auf die verdächtigen Übergangszonen aufmerksam machen, fungieren sie gleichsam als Brille. – Das letzte Wort hat freilich nie allein die Datenbank, sondern *wir* selbst, denn alles, was DA*mal*S bietet, sind objektive Messdaten, die zwar unser Gesamturteil bezüglich der Schreiberauthentizität wesentlich *plausibler* und vor allem intersubjektiv besser argumentierbar machen sollten, aber gemäß ihrer inhärenten ›fuzzy logic‹-Struktur letztlich nur Wahrscheinlichkeitswerte anbieten, welche es zu *interpretieren* gilt.

1.3 Komplementäres ›Eichungs-Szenario‹ in DA*mal*S

Die Bewusstheit für diese prinzipielle Interpretationsbedürftigkeit aller vergleichenden Messergebnisse in einem humanwissenschaftlichen Kontext hat schon am Projektbeginn neben das ›Authentifizierungs-Szenario‹, das primär der Ermittlung von allfälligen *Divergenzen* und daran knüpfbaren Schreiberhandgrenzen dient, ein ›Eichungs-Szenario‹ treten lassen. Dieses geht von einer bereits (durch eindeutige Quellenbeweise) identifizierten Schreiberhand aus und bemüht sich an ihrem Beispiel um das Ermitteln jener *Konvergenzen*, durch die – über Jahre und unterschiedliche Textsorten hinweg – eine Hand charakterisiert und wiedererkennbar scheint. Wie dazu in der bereits genannten DA*mal*S-Publikation (Hofmeister und Hofmeister-Winter) ausführlich dargestellt wird, haben wir dafür nach der *Bandbreite* des Individuellen im 15. Jahrhundert[11] gesucht, denn erst dadurch werden näherungsweise Urteile über die zu erwartenden Merkmals*schwankungen* bei noch nicht authentifizierten Händen möglich. Ausgewählt haben wir als ein solches ›Eichmaß‹ die umfangreichen (Ab-)Schriften der Lohnschreiberin Clara Hätzlerin. Noch konnten nicht alle ihre europaweit verstreuten Überlieferungsträger autopsiert und in unsere Datenbank eingespeist werden, aber doch die meisten. Dabei hat eine erste tentative Auswertung ergeben, dass sich anhand signifikanter Buchstaben- und ›Wortbilder‹ quer durch alle Stilisierungsebenen und Gebrauchshände dieser Augsburger Schreiberin in der Tat bereits so etwas wie ei-

[11] Und damit in zeitlicher Nähe der zu authentifizierenden Überlieferung der Werke Hugos von Montfort.

ne Schrift-›DNA‹ erkennen lässt, also rekurrente individuelle Schriftzüge. So hoffen wir (nach Abschluss unserer 3-Säulen-Befundung aller erhaltenen Schriftstücke der Clara Hätzlerin) jenen ›Maßstab‹ zu gewinnen, der – an noch nicht authentifizierte, aber zeitlich benachbarte deutschsprachige Hände angelegt – zumindest erahnen lässt, wo die Grenzen ›normaler‹ Ähnlichkeits-Streuungen liegen. Die künftige Einbeziehung weiterer identifizierter Schreiberhände, von denen es im Bereich deutschsprachiger Aufzeichnungen freilich speziell in spät- oder gar hochmittelalterlicher Zeit nur wenige gibt, könnte weitere ›Maßstäbe‹ beisteuern.[12]

2 ›Paläographische Schrifterfassung‹ (erste Authentifizierungs-Säule)

2.1 Allgemeines zum paläographischen Ansatz

Wie bereits oben erwähnt, wird der Paläographie im Methodenkanon von DA*mal*S zentrale Bedeutung beigemessen: Ihr kommt der erste prüfende Blick auf ein zu befundendes Schriftstück zu, um anhand formaler Merkmale eine grobe zeitliche Einordnung zu treffen; sie ist es auch, die bei der *Schlussbefundung* quasi das letzte Wort haben muss, wenn es gilt, die mit verschiedenen Methoden gewonnenen Einzelergebnisse in Relation zueinander zu setzen und ihre Plausibilität abzuwägen.

Die wesentliche Kompetenz der Paläographie als Hilfswissenschaft für alle historisch orientierten Disziplinen liegt unbestritten in der Hilfestellung bei der Entzifferung historischer Schriftsysteme, für die sie dank ihrer synchronen und diachronen Sammlungen von Schriftsymbol-Inventaren[13] sowie durch ›Schlüssel-Werke‹ aller Art[14] das notwendige Instrumentarium liefert. Erst damit wird es den Textwissenschaften möglich, verschriftlichte Sprache zu decodieren und so vielfältigen wissenschaftlichen Auswertungen zuzuführen.

Um diese Hilfestellung leisten zu können, muss die Paläographie den Blick auf die ›wesentlichen‹ Merkmale der Schrift einschränken, sie destilliert gleichsam die formalen Grundmuster einer Stilepoche aus einer Vielzahl von einzelnen Schriftzeugnissen durch Abstraktion von individuellen Ausprägungen. Genau auf diese individuellen Ausprägungen, die in mehr oder weniger deutlich erkennbaren *Abweichungen* von zeitlich bzw. regional gebräuchlichen Grundmustern bestehen, gilt es jedoch bei der Unterscheidung von Schreiberhänden zu fokussieren. Auch hier kann die Paläographie unterstützen, indem sie ein – wenn auch nicht allgemein verbindliches, so im Großen

[12] Vgl. die bemühte Dokumentation namentlich bekannter mittelalterlicher SchreiberInnen des europäischen Mittelalters durch Krämer.

[13] Zur Bestimmung des Alters und der Herkunft einer Handschrift steht eine Fülle von Tafelwerken und Übersichtstabellen zur Verfügung.

[14] Z. B. Cappelli, Grun.

und Ganzen doch brauchbares – *terminologisches Instrumentarium* zur Beschreibung des ›Augenbefundes‹ entwickelt hat.

Das Problem ist allerdings, dass die Beschreibung individueller Schriftzüge mit den Methoden der Paläographie – wie schon die oben erwähnte Aufstellung von ›Normen‹ für bestimmte Epochenabschnitte und Regionen – auf *Verallgemeinerung* des Befundes hinauslaufen muss: Die Suche nach den Charakteristika einer Schreiberhand führt zur Feststellung eines ›Durchschnitts‹ der beschreibbaren Merkmale, was gestützt auf paläographische Methoden zwar auf der Basis von Fachwissen und viel Erfahrung des Begutachters geschieht, sich aber letztlich doch zu einem Gutteil auf den *subjektiven* Eindruck gründet.

Um nun einerseits den paläographischen Befund in *intersubjektiv* nachvollziehbarer Weise festzuhalten, d. h. so *objektiv* wie möglich zu dokumentieren, und auf dieser Basis andererseits exakte Frequenzzählungen u.v.m. vornehmen zu können, ist es unumgänglich, den mit Hilfe der Paläographie decodierten Text zu ›recodieren‹ – idealerweise in Form einer *elementgetreuen Basistransliteration*, wie sie am Grazer Institut für Germanistik im Rahmen von Editionsaufgaben inzwischen zum Standard geworden ist und u. a. für den DA*mal*S-Beispielsfall Hugo von Montfort angewandt wurde.[15] Wie die Bezeichnung ›elementgetreu‹ bereits ahnen lässt, beschränkt sich diese Transliterationsmethode nicht auf die Umsetzung von alphabetischen Schriftsymbolen und allenfalls die Wiedergabe von buchstabenförmigen Superskripten und gängigen Abbreviatursymbolen im Stile diplomatischer Abdrucke, sondern sie verfeinert die Dokumentation zumindest so weit, dass sich darin *sämtliche* vertretenen Schriftelemente entsprechend codiert wiederfinden. In Details, die im Verdacht stehen, hinsichtlich der Individualität von Schreiberhänden Relevanz zu besitzen, kann die Transliteration sogar noch tiefer in Richtung einer ›phänomengetreuen Wiedergabe‹ gehen: Hier wird z. B. auch die *Form* von Superskripten speziell berücksichtigt, etwa auffällig variierende Häkchen-Formen. Aber auch den auf Abbildungen oft kaum sichtbaren feinen Haarstrichen, die von SchreiberInnen wohl weniger zur deutlicheren Differenzierung von Buchstabenformen (z. B. bei e, r, t) angebracht werden als aus ästhetischen Gründen, können charakteristisch für Schreiberindividuen sein und werden daher konsequent erfasst – und selbstverständlich später an den originalen Schriftstücken peinlich genau autopsiert. Dass bei den Superskripten konsequenterweise auch das Vorhandensein bzw. das Fehlen des i-Punkts extra verzeichnet wird, versteht sich nach den vorangegangenen Ausführungen fast von selbst.

In Form von *Annotationen* fließt in diese Basistransliteration auch Materielles ein: Mängel des (Be-)Schreibmaterials (Tinte, Papier/Pergament, Abnützungsspuren etc.)

[15] Entwickelt und erstmals angewandt wurde dieses Verfahren von Andrea Hofmeister-Winter anhand eines umfangreichen Editionsprojektes (Das Brixner Dommesnerbuch). Die ›elementgetreue Basistransliteration‹ bildet in diesem mehrstufig angelegten Editionskonzept die Grundlage für die gesamte weitere editorische Bearbeitung des Textes. Vgl. auch Hofmeister-Winter 2003 sowie ihre Forschungshomepage.

beeinflussen die Brauchbarkeit der Daten, so dass ›physisch beeinträchtigte‹ Belege ggf. aus dem Untersuchungsmaterial auszuscheiden sind. Das Hauptaugenmerk gilt aber dem ›Material‹ der Schriftzüge: Auch hier kann es vorkommen, dass Schriftsymbole durch Korrekturmaßnahmen (Tilgung/Rasur, Nachbesserung/Überschreibung) oder durch Schreiberversehen derart missgestaltet sind, dass sie nicht mehr als intakte Repräsentanten eines bestimmten Graphtyps anzusehen sind und daher für bestimmte Vermessungs- und Berechnungsoperationen nicht herangezogen werden können. Diesen Umständen wird durch entsprechende Annotationen Rechnung getragen.

Alle genannten Maßnahmen der objektiven Dokumentation des paläographischen Befundes fallen in den Bereich der ›niederen Textkritik‹ (Schieb), deren Aufgabenbereich nicht nur in der philologischen Editionswissenschaft oft zu wenig Beachtung findet, obwohl hier das Fundament für alle weiteren Analysen bis hin zur Interpretation im Rahmen der ›höheren Textkritik‹ im Sinne Karl Lachmanns und seiner AnhängerInnen gelegt wird (Hofmeister-Winter 2005). War die Herstellung einer deskriptiven Basistransliteration ursprünglich, d. h. im Rahmen einer ›dynamischen Edition‹ (Hofmeister-Winter 2003) in erster Linie als Hilfestellung für den Editor selbst gedacht, als eine Art ›*Wahrnehmungsprotokoll*‹ zur Schulung des editorischen Auges (Hofmeister 1999 33), erwies sich das solcherart gesicherte Informationsmaterial mittlerweile als vielfältig nutzbar: So konnten etwa die minuziös encodierten i-Punkte in den Schriften Veit Feichters[16] unsere ersten Auswertungen des Hugo von Montfort-Codex bestätigen und stützen (Hofmeister-Winter 2007 108–9).

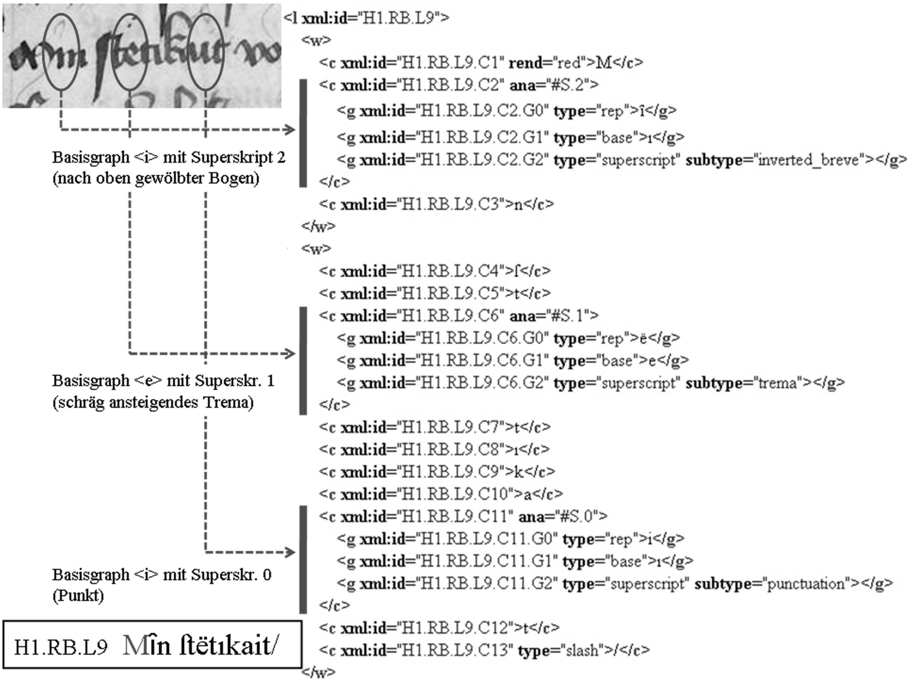
Zu den Prinzipien dieser ›*Mikro-Codierung*‹, die im Rahmen von DA*mal*S auf XML-Basis erfolgt,[17] gehört es, dass Superskripte getrennt von ihren Basisgraphen codiert werden. Auf diese Weise ist nicht nur die Kombinationsfähigkeit von Superskriptformen mit Basisgraphen besser analysierbar, sondern es können sog. ›verrutschte‹ Superskripte auf einer späteren Editionsstufe mit Hilfe ›höherer Textkritik‹ leichter korrigiert werden; auf der Stufe der Basistransliteration bewahrt diese Maßnahme davor, dass der paläographische Befund vorschnell durch emendierende oder gar konjizierende Interpretation ›verwischt‹ wird.

Welche Schriftsymbole in der Handschrift aufeinander bezogen sind und in welcher schreibräumlichen Relation das geschieht, ist durch die Codierung systematisch dokumentiert und kann daher aus der Transliteration (auch ohne Beiziehung der Handschriftenabbildungen) eindeutig rekonstruiert werden. In der XML-Transformation werden

[16] Außer dem »Dommesnerbuch« (Das Brixner Dommesnerbuch) sind von Veit Feichter auch ein »Urbar« und ein »Inventar« des Brixner Dommesneramtes erhalten; eine Edition auch der letzteren beiden Schriften durch Andrea Hofmeister-Winter ist in Vorbereitung.

[17] Bisher sind unsere Basistransliterationen im vertrauten Programm WinWord (als unformatierte ASCII-Dateien) entstanden und wurden von unseren Projektpartnern vom Zentrum für Informationsmodellierung in den Geisteswissenschaften (s. Anm. 8) für das Projekt DA*mal*S mit Hilfe von Transformationsregeln nachträglich in XML überführt; dieses Verfahren wird bis zur ›Marktreife‹ eines anwendungsfreundlichen Eingabetools (das zu den Nebenzielen des Projekts zählt) weiterhin beibehalten werden.

Basisgraphe und die dazugehörigen Superskripte als Glyphen (compound graphs) definiert, bestehend aus zwei diskreten Elementen, die in der Ausgabe-Anweisung beliebig (z. B. durch Unicodes) zur Darstellung gebracht werden können.[18] (Siehe Abb. 2: Beispiel für die XML-Codierung zusammengesetzter Zeichen.)


Basisgraph <i> mit Superskript 2
(nach oben gewölbter Bogen)
Basisgraph <e> mit Superskr. 1
(schräg ansteigendes Trema)
Basisgraph <i> mit Superskr. 0
(Punkt)
H1.RB.L9 Mîn ſtëtıkait/

```
<l xml:id="H1.RB.L9">
  <w>
    <c xml:id="H1.RB.L9.C1" rend="red">M</c>
    <c xml:id="H1.RB.L9.C2" ana="#S.2">
      <g xml:id="H1.RB.L9.C2.G0" type="rep">î</g>
      <g xml:id="H1.RB.L9.C2.G1" type="base">ı</g>
      <g xml:id="H1.RB.L9.C2.G2" type="superscript" subtype="inverted_breve"></g>
    </c>
    <c xml:id="H1.RB.L9.C3">n</c>
  </w>
  <w>
    <c xml:id="H1.RB.L9.C4">ſ</c>
    <c xml:id="H1.RB.L9.C5">t</c>
    <c xml:id="H1.RB.L9.C6" ana="#S.1">
      <g xml:id="H1.RB.L9.C6.G0" type="rep">ë</g>
      <g xml:id="H1.RB.L9.C6.G1" type="base">e</g>
      <g xml:id="H1.RB.L9.C6.G2" type="superscript" subtype="trema"></g>
    </c>
    <c xml:id="H1.RB.L9.C7">t</c>
    <c xml:id="H1.RB.L9.C8">ı</c>
    <c xml:id="H1.RB.L9.C9">k</c>
    <c xml:id="H1.RB.L9.C10">a</c>
    <c xml:id="H1.RB.L9.C11" ana="#S.0">
      <g xml:id="H1.RB.L9.C11.G0" type="rep">i</g>
      <g xml:id="H1.RB.L9.C11.G1" type="base">ı</g>
      <g xml:id="H1.RB.L9.C11.G2" type="superscript" subtype="punctuation"></g>
    </c>
    <c xml:id="H1.RB.L9.C12">t</c>
    <c xml:id="H1.RB.L9.C13" type="slash">/</c>
  </w>
```

Abbildung 2. Beispiel für die XML-Codierung zusammengesetzter Zeichen und die Darstellung in der ›Augenfassung‹ anhand Heidelberg, UB, cpg 329, fol. 1rb, Zeile 9.

Die elementgetreue Basistransliteration findet ihre obligatorische Ergänzung durch ein vollständiges Graphinventar, das einerseits einen Überblick über den in der Handschrift verwendeten Elementvorrat liefert und andererseits als *Transliterationsschlüssel* fungiert. Selbstredend gilt, dass das Verhältnis zwischen handschriftlichen Phänome-

[18] Zur sog. ›Augenfassung‹, einer leserfreundlichen Online-Synopse des cpg 329 und der elementgetreuen Basistransliteration, vgl. Wernfried Hofmeister: Perspektiven und Auswirkungen des Edierens am Beispiel der neuen Hugo von Montfort-Ausgabe [im Druck].

nen und Transliteration ›umkehrbar eindeutig‹ sein muss, damit der codierte Informationsgehalt 1 : 1 auf die handschriftliche Quelle rückführbar ist.[19]

2.2 Technische Umsetzung und Systemintegrierung der XML-Codierung

Für die Datenhaltung wurde – wie zuvor beschrieben – ein zweistufiger Ansatz gewählt: *imdas pro*, eine Applikation für die Verwaltung von Archivalien und Museumsobjekten, wird für die kodikologische Beschreibung und paläographische Befundung sowie für die Verwaltung der Schriftdokumente verwendet, wobei hier die *Digitalisate* mit eingespeist werden und so bereits in der Übersicht zur schnellen Orientierung zur Verfügung stehen. In der zweiten Stufe wird für jedes Dokument die Basistransliteration in einer TEI-konformen XML-Datei gespeichert. Die mächtigen Codierungsmöglichkeiten von TEI erlauben es, die Informationen zur Basistransliteration vollständig umzusetzen, wobei es im Speziellen aber als innovative Erweiterung dieser Möglichkeiten nötig war, für die Beschreibung von zusammengesetzten Buchstaben (Basisgraph und diakritisches Zeichen) eine Codierung über *compound-Zeichen* durchzuführen.[20] Das gestattete es in weiterer Folge, die einzelnen Komponenten getrennt zu markieren und darauf aufbauend die geometrischen Beziehungen zwischen Basisgraph und diakritischem Zeichen zu analysieren. Für jede Seite werden weiters (über ein ergänzendes *Tagging*) Angaben zur Position aller einzelnen Zeichen in einer eigenen Datei abgelegt,[21] wobei für jedes Zeichen auch die Referenz zum Zeichen in der TEI-Datei gespeichert wird. Die vom TEI-Dokument getrennte Speicherung dieser Informationen sowie die Auftrennung in einzelne Seiten ermöglicht ein gleichzeitiges, verteiltes Arbeiten.

Zur Erfassung dieser detaillierten Informationen wurde ein maßgeschneidertes Annotator-Tool umgesetzt, das die Eingabe der Basistransliteration erlaubt, dessen zentrale Aufgabe jedoch in der Erfassung der Zeichenpositionen liegt. (Das Userinterface des sog. DA*mal*S-*Annotators* ist in Abb. 3 zu sehen; dargestellt ist im oberen Teil ein Ausschnitt aus einer Handschriftenseite, in der ausgewählte Abschnitte annotiert wurden, im unteren Teil die dazugehörige Basistransliteration einschließlich der Zeilenzählung.) Zur Auswertung der erfassten Daten stellt der DA*mal*S-Annotator einerseits die Möglichkeit zur Verfügung, gezielt nach Zeichen, Buchstaben oder Buchstabenketten zu suchen und für die Fundstellen automatisch die entsprechenden Regionen aus den

[19] Vgl. die Hugo von Montfort-Editionshomepage (Anm. 9), wo der Transliterationsschlüssel den Basistransliterationen aller Überlieferungsträger unmittelbar beigeschlossen ist.

[20] An dieser Stelle sei Hubert Stigler und Petra Steinkellner vom Zentrum für Informationsmodellierung in den Geisteswissenschaften (vgl. Anm. 8) für ihre Entwicklung einer TEI-Codierung (mit speziellen Tags z. B. für Basisgraph-Diakritikum-Verbindungen) herzlichst gedankt.

[21] Diese Informationen sind in SVG codiert, je Zeichen wird ein Polygon (im einfachsten Fall ein Rechteck) gespeichert.

Digitalisaten auszuschneiden und gemeinsam mit der Positionsinformation als Einzelbilder abzuspeichern, womit – quasi als Abfallprodukt – der Aufbau eines graphischen Zeichen- bzw. Wortinventars möglich ist. Weiters sind die in Abschnitt 4.2 ausgeführten *Bildverarbeitungsmethoden* zur Unterstützung der Suche nach Schreiberhandwechseln – mit entsprechender visueller Darstellung – integriert.

3 ›Graphetische Statistik‹ (zweite Authentifizierungs-Säule)

3.1 Allgemeines zur graphetischen Frequenzanalyse

Statistische Verfahren der Korpusanalyse sind aus der modernen Linguistik nicht mehr wegzudenken: Sie erst führen durch die Möglichkeit der exakten *Quantifizierung* von rekurrenten Phänomenen zu einer Objektivierung des Befundes. Voraussetzung dafür ist die schon im Abschnitt 2.1 beschriebene Sicherung des Datenmaterials durch entsprechende Codierung: Welche Fakten zur Auswertung gelangen können, hängt maßgeblich von der ›Informationstiefe‹ des Datenmaterials ab. Speziell für die Untersuchung von Schreiberhänden ist es wichtig, dass die Transponierung einer Handschrift nicht erst auf der Graphemebene ansetzt, wo ein Großteil der paläographischen Beobachtungen bereits weggefiltert wurde, sondern möglichst nahe an der Handschrift, um den *maximalen Informationsgehalt* zu sichern.[22]

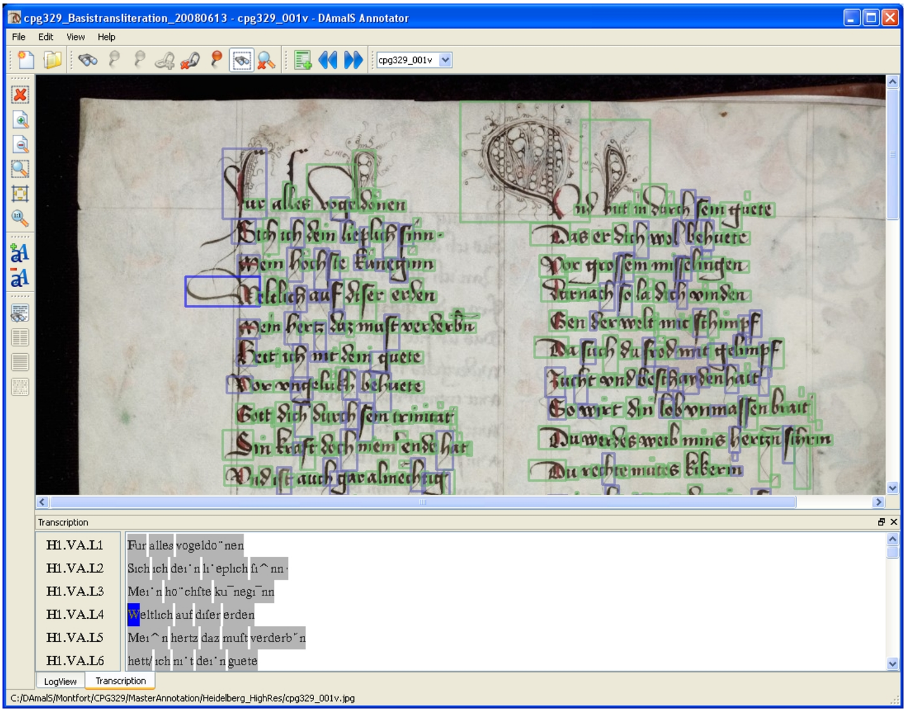
Der Mangel an solcherart aufbereitetem Material scheint ein plausibler Grund dafür zu sein, warum sich die statistische Methode mit ihrer exakten *Frequenzzählung* in der Paläographie bis heute nicht durchgesetzt hat; hier werden die ›Haupttendenzen‹ einer Hand, die aufgrund der Häufigkeit des Auftretens bestimmter Merkmale ins Auge stechen, nach wie vor durch relative und damit entsprechend unscharfe Frequenzangaben festgemacht.[23] Hemmen mag den Einsatz statistischer Methoden in der Paläographie auch die Befürchtung, dass z. B. die Berechnung von Durchschnittsmaßen von Buchstabenhöhe und -breite eher zur Nivellierung individueller Eigenarten führen könnte, anstatt diese sichtbar werden zu lassen (Schlögl 264–5).

Diese Gefahr besteht tatsächlich, wenn lediglich Mittelwerte für ganze Handschriften oder großräumige Abschnitte berechnet werden. Um Merkmalwechsel oder Brüche im Verlauf einer Handschrift erkennen und nachweisen zu können, ist es daher notwendig, die Berechnungen seiten- oder spaltenbezogen durchzuführen oder gar noch kleinere Zonen bis hin zu einzelnen Zeilen unter die Lupe zu nehmen. Es liegt auf der Hand, dass das nur mittels solcher Merkmale sinnvoll ist, die entsprechend häufig vorkommen. Daher konzentriert sich DA*mal*S eben nicht in erster Linie auf seltene, auffällige Merkmale, sondern auf Buchstabenformen, -kombinationen bis hin zu Silben und Wörtern,

[22] Vgl. Hofmeister-Winter 2005 6 (Graphische Darstellung der Relation von Befund und Deutung).

[23] Vgl. z. B. die jüngste Arbeit von Schneider 2007, die rein auf relativen Frequenzangaben basiert; erste Ansätze von statistischen Methoden in der Paläographie referiert Bromm.

die möglichst zahlreich aufscheinen und einigermaßen homogen über den gesamten Text gestreut sind (vgl. Abschnitt 4.1).


cpg329_Basistransliteration_20080613 - cpg329_001v - DAmalS Annotator
File Edit View Help
cpg329_001v
Transcription
LogView Transcription
C:/DAmalS/Montfort/CPG329/MasterAnnotation/Heidelberg_HighRes/cpg329_001v.jpg

H1.VA.L1	Fur alles vogeldöͤnen
H1.VA.L2	Sıch ıch deı̇n lı̇eplıch ſı̂nn ·
H1.VA.L3	Meı̇n höchſte kūnegı̄nn
H1.VA.L4	Weltlıch auf dıſer erden
H1.VA.L5	Meı̂n hertz daz muſt verderb̆n
H1.VA.L6	hett/ ıch nı̇t deı̇n guete

Abbildung 3. DAmalS-Annotator Screenshot.

Welcher Erkenntnisfortschritt auf der Materialbasis einer elementgetreuen Basistransliteration zu erzielen ist, sei im Folgenden anhand der Heidelberger Montfort-Handschrift cpg 329 exemplarisch vorgeführt: Bereits 1881 hatte Josef Wackernell in Zusammenhang mit seiner Edition der Texte des adligen Dichters (Hugo von Montfort 1881) ausführliche Untersuchungen zu den beteiligten Schreiberhänden angestellt, mit denen er sich als ›Kronzeuge‹ des gesamten interdisziplinären Methodenrepertoires seiner Zeit erwies und ein großartiges Lehrstück der mediävistischen Textforschung lieferte. Seine verdienstvollen paläographischen und sprachwissenschaftlichen Befundungen bildeten einerseits die Ausgangsbasis, andererseits eine Art ›Reibebaum‹ für die Analysen des Codex im Rahmen des Projekts DA*mal*S. Wackernells Ergebnisse galt es in einem ersten Schritt auf ihre Stichhaltigkeit zu überprüfen und – da sein Resultat nicht restlos überzeugte – durch weitere Untersuchungen zu ergänzen.

Wackernell ging bei seinen Untersuchungen systematisch ›von außen nach innen‹ vor und bemühte sich sichtlich um Glaubwürdigkeit seiner Behauptungen, indem er (im Rahmen seiner bescheidenen technischen Möglichkeiten) Häufigkeitsangaben machte. Frequenzzählungen gestalteten sich in den Anfängen der Sprachwissenschaft denkbar mühsam, konnten sie doch lediglich auf der Basis händischer Auszählung erfolgen. Absolute Frequenzzählung findet sich daher nur bei selten auftretenden Phänomenen, während ansonsten relative Häufigkeitsangaben bevorzugt werden: ›stets – meist – häufig – selten – nie‹, so lautet in etwa die gängige Skala dieser intuitiven Befundungsmethode. Selbst wenn gelegentlich konkrete Zahlenangaben angeführt werden, sind diese oft nicht nachvollziehbar – nicht so sehr aufgrund von Irrtum (der gerade in diesem Bereich im wahrsten Sinn des Sprichworts *Errare humanum est* verzeihlich wäre), sondern weil nicht immer deutlich genug deklariert ist, nach welchen Kriterien Belege gezählt oder ausgeschieden wurden (Hofmeister-Winter 2007 94–5).

Als Erstes begutachtete Wackernell das Layout des Codex und stellte Diskontinuitäten beim Seitenspiegel resp. bei der Schreibraumeinteilung fest, die sich konkret an der durchschnittlichen Zeilenzahl pro Spalte festmachen ließen und ihre Ursache in einer geringfügig abweichenden Schriftgröße haben. Bei der paläographischen Schriftanalyse konzentrierte sich Wackernell auf Einzelmerkmale wie z. B. die Schaftform von Lang-s, f und p oder bestimmte Verzierungen an Ober- bzw. Unterlängen der ersten/letzten Zeile einer Seite sowie auf das Auftreten von Graphvarianten wie etwa Ligatur-r. Im Bereich der Graphie-Unterschiede untersuchte Wackernell die Verwendung von Majuskel <R>, die (in Text Nr. 38) nicht nur obligatorisch am Zeilenanfang, sondern auch im Zeileninneren auftritt. Graphieunterschiede boten sich aber auch im Bereich der graphischen Umsetzung bestimmter Phoneme für eine genauere Untersuchung an; so erscheint z. B. die Wiedergabe des frühneuhochdeutschen Diphthongs [ei] in den von Wackernell ausgemachten Zonen teils in den moderneren Diphthong-Graphien <ei, ey, ai, ay>, teils in der konservativen (weil noch nicht diphthongierten) Schreibung mittelhochdeutscher Handschriften als <i> (ggf. mit Superskript).

Aus der Summe dieser und weiterer Beobachtungen zog Wackernell den Schluss, dass an der Niederschrift des cpg 329 insgesamt vier Schreiberhände beteiligt gewesen seien, denen er mangels Identifizierbarkeit die Bezeichnungen A, B, C und D zuwies. Er legte auch die Zonen fest: Demnach habe Hand A fol. 1r–12v geschrieben (diese Grenze ist zugleich eine Lagengrenze, wenn sie auch mitten durch Text Nr. 12 verläuft), Hand B fol. 13r–46v, Hand C fol. 47r–48va und Hand D (als einzige deutlich und zweifelsfrei unterscheidbar von den übrigen angenommenen Händen) fol. 48vb–52va (Hugo von Montfort 1881 CXII–CXX; vgl. Hofmeister-Winter 2007 93, Tab. 1).

Die von Wackernell sorgfältig gesammelten Indizien zeigen leider einen Mangel: Sie sind großteils nicht wirklich tragfähig – zu wenig ist über die Entstehungsumstände des Codex, die Qualität (Homogenität/Heterogenität) der Vorlage(n) und folglich über den

Umgang der von Hugo beauftragten Kopisten mit ihren Vorlagen bekannt.[24] Daher lassen sich zu allen Hypothesen Wackernells Zweifel und Gegenargumente anführen, was die Forschung der jüngeren Vergangenheit zu einem radikalen Rückzug veranlasste: Einige gingen sogar so weit, für die drei fraglichen Zonen A, B und C mangels ausreichender Beweise überhaupt nur mehr einen einzigen Schreiber anzunehmen (Spechtler; Werner; Welker).

Auch an der methodischen Vorgehensweise Wackernells gibt es aus Sicht der modernen Linguistik einiges zu kritisieren: Pauschalaussagen über paläographische Detailbeobachtungen haben höchstens den Status persönlicher Eindrücke, wenn sie nicht objektiv belegt werden können[25] – man bedenke, dass die zeitgenössischen Leser seiner detaillierten Abhandlung über die Schreiber des cpg 329 noch nicht einmal ein Faksimile zur Verfügung hatten, um die Aussagen wenigstens visuell zu prüfen. Eine Möglichkeit, hier Abhilfe zu schaffen, bestünde in der (partiellen) Vertiefung der Transliteration, wie sie am Beispiel der <h>-Varianten demonstriert werden konnte. Um auch die *mikroskopischen* Feinheiten der Graphvarianz zu erfassen, wurde von uns eine *Komponentenanalyse* versucht, die den Graph in seine morphologischen Bauelemente (Schleife, Schaftfuß und Bogen) zerlegt und diese Variablen jeweils getrennt beschreibt (Hofmeister-Winter 2007 97–100, bes. Tab. 2, und 114, Abb. 9). Diese Untersuchung, die sich wegen des nicht unerheblichen Deskriptionsaufwandes auf die ›neuralgischen‹ Zonen des Codex (rund um die von Wackernell konstatierten Schreibergrenzen) beschränkte, lieferte tatsächlich signifikante Ergebnisse hinsichtlich der von A, B und C bevorzugten Kombinationsformen, zeigte aber auch deutlich das Problem der *stilistischen Schwankungsbreite* von Schreiberhänden auf, das von der Forschung bisher noch nicht eingehend untersucht worden ist.

Problematisch an Wackernells Vorgehen ist weiters die Beiziehung von Phänomenen mit sehr geringer Belegzahl oder gar ›Hapaxlegomena‹: Es stimmt zwar, dass im cpg 329 die Graphvariante Ligatur-r in der angenommenen Zone A überhaupt nicht vorkommt, aber da sie in Zone B und C nur 1% aller <r>-Belege stellt, ist die Signifikanz dieses Phänomens als äußerst schwach einzustufen (Hofmeister-Winter 2007 100–1). Wackernell findet sich zwar in Einklang mit der traditionellen Paläographie, die bei der Beurteilung von *Schreiberindividualität* seit jeher den Blick schwerpunktmäßig den Abweichungen, dem Außergewöhnlichen widmet, das eben oftmals nur vereinzelt auftritt. Jedoch er-

[24] Im Fall der Texte Hugos von Montfort ist damit zu rechnen, dass diese ursprünglich der vorarlbergischen Herkunft des Dichters entsprechend mehr oder weniger stark alemannisch gefärbt waren; die Eintragung in den von Hugo selbst in Auftrag gegebenen Codex erfolgte jedoch in der Zeit, die er aus privaten wie beruflichen Gründen in der Steiermark verbrachte (ab ca. 1414), vermutlich durch Schreiber aus der Region, welche die Texte womöglich eigenständig in unterschiedlichem Maße bairisch überformten (wobei ungeklärt bleiben muss, wie weit nicht schon die Kopiervorlagen der über mehrere Jahrzehnte entstandenen Dichtung dialektal schwankten). Vgl. Hugo von Montfort 2005 XXIV-XXV.

[25] Vgl. z. B. Wackernell über die Schaftform von Lang-s bei den verschiedenen Händen (Hugo von Montfort 1881 CXVII).

schweren selten vertretene Merkmale die Bestimmung von Schreibergrenzen insofern, als keine flächendeckende Markierung der Schreiberbereiche gegeben ist. Schon aus diesem Grund sollte das Augenmerk – nicht nur, aber mit gleich viel Akribie – auf *hochfrequente* Phänomene gelegt werden, die eine möglichst gleichmäßige Streuung über den ganzen Text aufweisen.

Sowohl Layout- und schriftstilistische als auch Graphieunterschiede dürfen nicht vorbehaltlos als Ausdruck von Schreiberindividualität betrachtet werden. Sie stehen unter dem Einfluss vielfältiger Faktoren, die oft außerhalb des Schreiberindividuums liegen und daher nicht von diesem gesteuert werden; entsprechend umsichtig gilt es, sie im Rahmen von statistischen Auswertungen zu berücksichtigen:

- Das *Layout* kann z. B. vom Auftraggeber veranlasst sein, der sich ein bestimmtes Design wünscht; im Fall von Arbeitsteilung müssen sich alle Schreiber dem Gesamtkonzept beugen, dies gilt vor allem ab dem zweiten Schreiber. Dennoch sind Umstände denkbar, die dazu führen, dass die Kontinuität des Layouts durchbrochen wird: Ein Wechsel der Schreiberhand ist nur eine Möglichkeit unter vielen.
- Die abwechselnde/gleichzeitige Verwendung verschiedener *Formvarianten* eines Schriftsymbols ist noch kein Beweismittel für die Authentifizierung einer Schreiberhand, denn es ist nicht unüblich, dass ein (zumal geübter) Schreiber mehrere stilistische Varianten in seinem Repertoire hat, ja sogar mehrere Duktus (sog. Anlasshände), die er entweder willkürlich, kontextabhängig oder aber bewusst für verschiedene Zwecke/stilistische Ansprüche einsetzt.[26]
- Unterschiede der *Graphie* schließlich, also Abweichungen hinsichtlich der Umsetzung gesprochener Laute in Schrift, können, müssen aber nicht auf den individuellen Schreibusus der beteiligten Personen zurückgehen, sodass diese den Vorlagentext sozusagen bewusst ihren eigenen Graphiegewohnheiten und damit indirekt ihrem Dialekt angeglichen hätten. Es muss immer auch mit inhomogenen Vorlagen gerechnet werden.

Die Phänomene, in denen Wackernell (und mit bzw. nach ihm die traditionelle Paläographie) auch anhand statistischer Gesichtspunkte nach Schriftindividualität gesucht hat, hängen großteils von äußeren Einflüssen ab. Für die Fahndung nach Schriftindividualität sind jedoch besser solche *Schriftmerkmale* aufzuspüren und auszuwerten, die möglichst ›*unbelastet*‹ von Vorbildern und Normen sind und am besten beim Schreibvorgang nicht im Zentrum der Aufmerksamkeit des Schreibers liegen und daher nicht ›bewusst‹ ausgeführt werden. Derartige Merkmale sind eher im peripheren Bereich des Schriftsystems zu entdecken und in der Regel dadurch charakterisiert, dass sie nicht un-

[26] So beobachtet bei der Augsburger Berufsschreiberin Clara Hätzlerin, die u. a. zwei völlig verschiedene <d>-Formen gebraucht, die sie teils (fast) ausschließlich, teils willkürlich vermischt einsetzt. Vgl. Hofmeister und Hofmeister-Winter 2008 104–7 und 114–5 (Abb. 4–6).

mittelbar sprachsystemrelevant sind, sondern redundant oder als Zierelemente ›überflüssig‹ und daher in gewisser Weise entbehrlich (Hofmeister-Winter 2007 106–7).

Zwei Elemente der Schrift, auf die diese Eigenschaften in hohem Maß und dabei mit statistisch signifikanter Frequenzstreuung zutreffen, scheinen der *i-Punkt* und haarfeine *Zierstriche* an bestimmten Graphen zu sein: Diese unscheinbarsten Elemente des Schriftsystems, deren Existenz sich oft nur am Original verifizieren lässt, werden aus schreibpsychologischer Sicht kaum jemals bewusst gesetzt, sondern verdanken sich als periphere Merkmale der betreffenden Schriftsymbole eher einem *graphomotorischen Reflex*, ausgelöst durch den einmal erworbenen Usus. Die i-Punkt-Setzung scheint – zumindest in der Zeit von ihrem ersten Auftreten ab 1320 (Schneider 1999 49) bis zur Normierung des i-Punkts im Schriftsystem als ›Diakritikum‹ – deshalb noch in starkem Maße ›unbelastet‹, weil sie dem Belieben des Individuums unterstellt und daher weitgehend vorlagenunabhängig ist. Für gegenwartssprachliche handschriftliche Texte liegen noch keine Vergleichszahlen vor, jedoch sollte sich hier die Frequenz aufgrund der Normiertheit des i-Punktes der 100 %-Marke annähern. Dennoch ist denkbar, dass die durchschnittliche Zahl der ›vergessenen‹ i-Punkte auch in der Gegenwart Schreiberindividuen charakterisiert. Solche Schriftelemente können als *psychometrische Merkmale* bezeichnet werden, da sie den Schriftzügen quasi wie eine DNA eingeschrieben sind und in Summe die ›Grammatik der Schreiberhände‹ prägen. In der Basistransliteration des cpg 329 wurden einige dieser Schreibereigenheiten mittels der oben beschriebenen Codierungsmethode festgehalten (vgl. Abschnitt 2.1).

Zur statistischen Beobachtung von charakteristischen Zügen einer Hand im Bereich der schriftsystemrelevanten Elemente eignen sich ganz besonders auch *Abbreviaturen*. Wieder gilt es zunächst das Repertoire zu erheben und in der Basistransliteration differenziert zu codieren, sodann Frequenz und Streuung der verschiedenen Formtypen zu ermitteln und schließlich mit Hilfe einer graphetischen Analyse die oft individuell geprägte Verwendungsweise für bestimmte Graphsequenzen oder formelhafte Wörter zu untersuchen. Zusätzlich kann die Ermittlung des generellen ›*Kürzungsindikators*‹ lohnend sein, also der Frequenz des Einsatzes von Kürzungssymbolen im Allgemeinen (am besten zeilenbezogen) – dies aber nur unter bestimmten Voraussetzungen, nämlich wenn ein Codex strikte Layoutvorgaben mit Randausgleich (in der Typographie als ›Blocksatz‹ bezeichnet) aufweist, der die Schreiber zur ökonomischen Befüllung des zur Verfügung stehenden Schreibraumes zwingt.[27] Das ist im cpg 329 allerdings nicht der Fall, weshalb keine Notwendigkeit zu Kürzungen bestand, so dass Schriftelemente dieser Kategorie eher sporadisch auftreten und ihre Frequenzwerte geringe Aussagekraft besitzen.

[27] Vgl. Hofmeister 2001 94. Das Ziel dieser Studie war allerdings etwas anders gelagert: Hier ging es darum, im cpg 848 verschiedene Schreibschichten nachzuweisen, nämlich Grundschicht und Nachträge, für die von den Grundschichtschreibern ein exakt bemessener Freiraum ausgespart worden war, der von den Nachtragsschreibern möglichst ohne erkennbare Bruchlinie ausgefüllt werden musste.

Die statistische Auswertung des ›mikro-codierten‹ Materials kann beispielsweise folgende Fragestellungen behandeln:

- Ob z. B. i-Punkte gesetzt sind bzw. in welcher *durchschnittlichen Frequenz*: Schon diese Werte können von Schreiberindividuum zu Schreiberindividuum signifikant differieren. Für eine objektive Feststellung des Durchschnitts ist es allerdings erforderlich, von den durch Wackernell prädisponierten Zonen abzusehen und das Material in einer Weise zu analysieren, die Frequenzschwankungen unvoreingenommen anzeigt. Wir haben zu diesem Zweck den Wert für jede Spalte gesondert ermittelt, und zwar die Belegzahl pro Zeile. Im cpg 329 kristallisierten sich auf diese Weise drei deutlich voneinander abgesetzte Zonen heraus: A, B/C und D (Hofmeister-Winter 2007 107 und 116, Abb. 14). Im Vergleichskorpus von Veit Feichter (ausgewertet wurden alle drei bekannten Codizes des Schreibers) zeigte die durchschnittliche Frequenz des i-Punktes innerhalb von gut zehn Jahren lediglich minimale Abweichungen.[28]
- Ob der i-Punkt-Gebrauch bestimmten *Normen* folgt: Hier ergaben sich im Fall der Analyse der Schriften der Clara Hätzlerin, die dem Projekt DA*mal*S als ›Eichwerkzeug‹ dient (vgl. Abschnitt 1.3), erstaunliche Einsichten, nämlich dass diese routinierte Schreiberin offenbar für sich klare Regeln entwickelte, die sie noch dazu peinlich genau einhielt. Die genauere Untersuchung dieser Regeln ergab, dass die Schreiberin äußerst ökonomisch verfuhr, indem sie i-Punkte nur dort setzte, wo sie zur Differenzierung des Basisgraphs <i> unbedingt nötig sind bzw. den Lese-/Verstehensprozess unmittelbar fördern, also in direkter Nachbarschaft mit solchen Buchstaben, die aufgrund ihrer Bauweise (Kurzschäfte) zu Missverständnissen führen können; das sind im Fall des kursiven Schreibsystems der Hätzlerin u, m, n (Hofmeister und Hofmeister-Winter 107–8).

Die Eruierung dieser schreiberindividuellen Normen liefert viel exaktere Befunde als die bloße Ermittlung der durchschnittlichen i-Punkt-Frequenz innerhalb eines Textes. Da eine verbindliche überindividuelle Schreibnorm im 15. Jh. noch nicht ausgeprägt war, kann die Erstellung eines *Schreiberprofils* wertvolle Indikatoren für die Authentifizierung von Schreiberhänden liefern. Die Zuverlässigkeit dieser Methode muss allerdings durch die serienmäßige Analyse anderer Schreiber weiter abgesichert werden.

Durchgeführt wurden die beschriebenen Untersuchungen mit Hilfe von Beleglisten, die zu jedem Beleg auch den genauen Fundort angeben, so dass jeder einzelne Repräsentant eines bestimmten Phänomens bis in die Handschrift zurückverfolgt werden kann. Dies ist zur Kontrolle der Belege unerlässlich, denn da Transliterationen in der Regel nicht von allem Anfang an die Disambiguierung aller einzelnen Wortformen durch Annotationen vorsehen, müssen Homographen auf diese Weise identifiziert und ggf.

[28] Die Abweichung vom Gesamtdurchschnitt (98,8 %) beträgt weniger als 1 %. Vgl. Hofmeister-Winter 2007 108–9.

ausgesondert werden. Zugleich bietet die *exakte Beleg-Verortung* maximale Transparenz

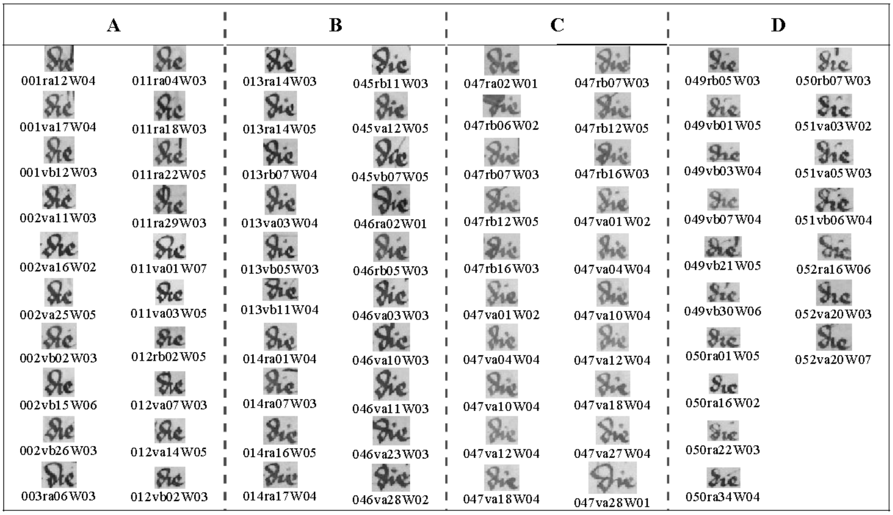

A
B
C
D
001ra12W04
011ra04W03
013ra14W03
045rb11W03
047ra02W01
047rb07W03
049rb05W03
050rb07W03
001va17W04
011ra18W03
013ra14W05
045va12W05
047rb06W02
047rb12W05
049vb01W05
051va03W02
001vb12W03
011ra22W05
013rb07W04
045vb07W05
047rb07W03
047rb16W03
049vb03W04
051va05W03
002va11W03
011ra29W03
013va03W04
046ra02W01
047rb12W05
047va01W02
049vb07W04
051vb06W04
002va16W02
011va01W07
013vb05W03
046rb05W03
047rb16W03
047va04W04
049vb21W05
052ra16W06
002va25W05
011va03W05
013vb11W04
046va03W03
047va01W02
047va10W04
049vb30W06
052va20W03
002vb02W03
012rb02W05
014ra01W04
046va10W03
047va04W04
047va12W04
050ra01W05
052va20W07
002vb15W06
012va07W03
014ra07W03
046va11W03
047va10W04
047va18W04
050ra16W02
002vb26W03
012va14W05
014ra16W05
046va23W03
047va12W04
047va27W04
050ra22W03
003ra06W03
012vb02W03
014ra17W04
046va28W02
047va18W04
047va28W01
050ra34W04

Abbildung 4. Liste von <die>-Belegen mit exakten Positionsangaben.

und erleichtert die Nachvollziehbarkeit der Untersuchungen (vgl. Abb. 4; die Stellenangabe umfasst Blattzahl, Spalte, Zeile sowie – mit der Sigle »W« – die Wortnummer des Belegs innerhalb der Zeile).

3.2 Technische Umsetzung der statistischen Auswertung

Ausgehend von der im Abschnitt 2.2 dargestellten detaillierten Erfassung aller Schriftsymbole und Zeichenpositionen ergibt sich eine Vielzahl von Möglichkeiten, diese Grundinformationen statistisch auszuwerten. So kann man z. B. automatisch Zeichen- und Wortindizes – ergänzt durch entsprechende Belege aus dem Digitalisat – erzeugen oder Zeichen- und Worthäufigkeiten für den gesamten Codex, je nach Fragestellung auch pro Seite, Spalte oder Zeile, evtl. zusätzlich hinsichtlich Buchstabenvarianten detailliert, berechnen. Für weitergehende Auswertungen und die Erzeugung entsprechender Schaubilder ist der Export über den DA*mal*S-Annotator in standardisiertem Format (csv-Dateien) vorgesehen: Damit lassen sich dann (auch via Excel-Dateien) *Visualisie-*

rungen z. B. der Frequenzverteilung anhand besonders aussagekräftiger graphischer Diagramme generieren.[29]

4 ›Musterorientierte Schriftbild-Erfassung‹ (dritte Authentifizierungs-Säule)

4.1 Allgemeines zum musterorientierten Ansatz

Jegliche mündliche Sprachreproduktion bedient sich eingeübter Muster, indem nicht Einzellaute artikuliert werden, sondern miteinander zu Klangbildern verschmelzende Laut*folgen* von Silben oder Wörtern; als individuell eingeübte Schemata steigern sie die Ökonomie und Persönlichkeit unserer Kommunikation. Wenn Sprache *schriftlich* ausgedrückt wird und auch diese ›sekundäre‹ Form der Sprachreproduktion flüssig (also nicht Buchstaben für Buchstaben malend, sondern etwa in routinierter Kursive) erfolgt, ist mit demselben Phänomen einer ›Amalgamierung‹ zu rechnen:[30] Dies mag man sich dadurch erklären, dass die Art, wie die Buchstaben innerhalb einzelner Silben oder Wörter zusammengefügt werden, ebenfalls einer Art *mentaler Matrix* gehorcht, durch deren eintrainiertes ›Lexikon‹ die Feinmotorik einer schreibenden Hand gesteuert wird; ähnlich wie auf der Laut-Ebene ist dann auch in den Schriftzügen mit charakteristischen Verschleifungen und ensembleartigen *Musterbildungen* zu rechnen. Dies darf zum einen für komplexer geformte Buchstaben und Ligaturen angenommen werden, zum andern aber auch für ganze Silben bzw. Morpheme und Wörter: Bei deren Verschriftung kann es daher zu *stempelartigen* Ausprägungen kommen.

Die ›klassische‹ Paläographie hat ihre bisherigen Befundungen von Schriftzügen fast ausschließlich auf die Beobachtung von Einzelbuchstaben gestützt, allenfalls ergänzt um Ligaturen. Damit hat sie unbestreitbar ein beachtliches Register an Kriterien entwickelt, das außer für die *zeitliche* und *räumliche* Einordnung von Schriftproben auch für das Erkennen von *individuellen* Schrift-Charakteristika und damit für das Unterscheiden einander recht ähnlicher Handschriften äußerst hilfreich ist (Schneider 1999). Dennoch blieben – wie schon am Beginn dieses Beitrags festgestellt – zahlreiche Zweifelsfälle bislang ungeklärt. Um nun sowohl auf dieser graphetischen Ebene als auch auf der (hier vergröbert so genannten) ›Wort‹-Ebene eine Verbesserung unserer Seh- und Trennschärfe zu erreichen, bedient sich DA*mal*S aller modernen Möglichkeiten der *digitalen Bildverarbeitung*, indem elektronische Abbildungen von Handschriften mit einbezogen werden.

[29] Diese Funktionalitäten werden vom DA*mal*S-Annotator derzeit nur teilweise angeboten, sollen jedoch im Fortsetzungsprojekt MOSES komplettiert werden.

[30] Zur gestalthaften Abbildung lexikalischen Sprachmaterials vgl. Aitchison. Über allgemeine graphematische Zusammenhänge reflektiert anschaulich Grabowski.

Voraussetzung für eine graph- und wortgenaue Befundung von Schriftzügen ist die Verknüpfung aller Transliterations-Elemente mit den dazugehörigen Bildinhalten. Sie wird (vom unten näher beschriebenen und eigens für DA*mal*S entwickelten) *Annotator-Tool* geleistet: In zwar sehr mühsamer, am Ende jedoch ebenso lohnender Arbeit werden mit seiner Hilfe alle Graphe und Graph-Elemente durch eine exakte Positionsangabe mit dem Bild verbunden. Erst durch diese individuelle *Adressierung* gelingt es, von Beleg zu Beleg Vergleichsmessungen durchzuführen und entsprechende Divergenz- oder Konvergenzwerte zu gewinnen, die in die XML-Beschreibung einfließen können.

Um nun etwa für diverse *Leitbuchstaben/Kenngraphien* vergleichbare Messwerte zu erhalten, werden entweder ganze Buchstaben oder auch nur auffallend variante Teile von ihnen (wie Unterlängen, Schlaufenformen oder Diakritika) herausgegriffen und miteinander verglichen. Die bislang überzeugendsten Erfolge können dabei (wie schon oben erwähnt) durch die gesonderte Erfassung der Positionierung des i-Punkts erzielt werden, weil dieses Auszeichnungselement zu jenen gehört, die sich allem Anschein nach einer bewussten Steuerung entziehen und entsprechend viel über die individuelle *psychomotorische Routine* einer Schreiberhand verraten.

Darüber hinaus wird der Blick aber auch auf die bislang fast vollständig ignorierte *Wort-/Silben*-Ebene erweitert; auch dafür bietet die elektronische Bildverarbeitung bislang ungeahnte Möglichkeiten: So können alle dafür in Frage kommenden Graphsequenzen als ganze vermessen, ausgeschnitten und in diverse Listen exportiert werden, wo sie dann – nach Ähnlichkeit oder Position dicht an dicht gereiht – in Form eines ›virtuellen Morphings‹ neu beurteilbar werden. Am Beispiel des Wortes »die« scheint dies durch DA*mal*S bereits geglückt, und zwar sowohl für die zu authentifizierenden Hände im cpg 329 (vgl. Abb. 4) als auch für die identifizierte Hand der Clara Hätzlerin (Hofmeister und Hofmeister-Winter).

Die soeben erwähnte Auswahl des bestimmten Artikels »die« als ›Stempelwort‹ führt zur Frage nach ihrem methodischen Hintergrund und einer Ausschau nach weiteren tauglichen Wörtern für eine musterorientierte Untersuchung. Diese Selektion gehorcht, kurz gesagt, folgenden Kriterien: Als Stempelwort eignen sich lexikalisch-morphologische oder silbische Einheiten, sofern sie *hochfrequent* sind und dazu – in Summe – *alle paläographisch wichtigen Kenngraphien* abdecken. Moderne Wortfrequenzlisten helfen zwar bei der Einengung dieser Suche in historischen Texten,[31] da eben Grundformen aus dem Bereich der Pronomina, Präpositionen oder Konjunktionen seit jeher häufig auftreten, aber eine darauf fußende Liste wird nicht für alle An-

[31] Vgl. die Statistik des Projekts »Wortschatz« des Instituts für Informatik/Abteilung für Sprachverarbeitung an der Universität Leipzig [zuletzt eingesehen am 7.2.2009]. Demnach sind die 25 häufigsten Wortformen der deutschen Gegenwartssprache (in absteigender Reihenfolge): der, die, und, in, den, von, zu, das, mit, sich, des, auf, für, ist, im, dem, nicht, ein, Die, eine, als, auch, es, an, werden. Eine ähnliche Statistik lässt sich auch für diverse frühneuhochdeutsche Textcorpora ermitteln (vgl. die folgende Anm.).

wendungsfälle geeignet sein,[32] denn da gilt es auf textsortenspezifische Bedingungen Rücksicht zu nehmen: Durch derartige Grundperspektiven von Texten kann z. B. das Auftreten des an sich häufigen Personalpronomens »ich« ausgeschlossen oder stark eingeschränkt sein. Somit empfiehlt es sich, ein *Register* von präsumtiven Stempelwörtern/musterbildenden Graphsequenzen anzulegen, aus dem dann für jeden Untersuchungsfall ein *maßgeschneidertes Bündel* (mit vielen Leitgraphien) ausgewählt wird.

In Summe können alle erwähnten Prozeduren für die schon mehrfach erwähnte *Brillenfunktion* von DA*mal*S genützt werden: Durch Visualisierung aller Messdaten, die sich (entlang der Aufzeichnungsrichtung) als Abweichungen von bis dahin etablierten Stempelwort-Grundmustern ergeben haben, kann man jede signifikante Abweichung direkt im digitalen Faksimile markieren. Wie sich diese ›*Rückprojektion*‹ technisch umsetzen lässt, wird nun in Abschnitt 4.2 erläutert.

4.2 Technische Umsetzung der Mustererkennung

Um das Ziel von DA*mal*S – eine intersubjektiv besser argumentierbare Entscheidung hinsichtlich des Auftretens eines Schreiberwechsels – zu unterstützen, wurden Bildverarbeitungsmethoden auf der *Makroebene* (Einbeziehung des gesamten Codex bzw. zumindest mehrerer Seiten oder Spalten daraus) und der *Mikroebene* (Untersuchung und Vergleich von Zeichen, Buchstaben oder Wörtern) umgesetzt. Das Ergebnis der automatisch ablaufenden Methoden der Makroebene ist die unmittelbare – mit Unsicherheit behaftete – Einschätzung, ob ein Schreiberwechsel aufgetreten ist. Die Mikroebene ergibt weitere objektive Informationen, die den subjektiven Entscheidungsprozess unterstützen.

Der Entscheidungsprozess auf der *Makroebene* läuft wie folgt ab:[33] In einem ersten Schritt werden in den Digitalisaten jene Bereiche identifiziert, die Text enthalten, da die nachfolgenden Analysen durch graphische Elemente (Fleuronnée etc.) gestört werden würden. Danach werden diese Textblöcke in Segmente (die ein oder mehrere Zeilen enthalten können) getrennt und diese ihrerseits in Regionen zerlegt, die grob Buchstaben gleichzusetzen sind. Diese Regionen werden nun weiter analysiert, wobei einerseits Maße errechnet werden, die von der Kontur der Region abgeleitet werden, andererseits solche, die auf der Strichstärke in bzw. Abständen zwischen Regionen beruhen.

Basierend auf der Kontur werden drei Maße berechnet: der Winkel eines Kontursegmentes gegenüber der Horizontalen, der Winkel, den zwei aufeinanderfolgende Kon-

[32] Ein hilfreiches Tool zur Frequenzbestimmung von Wörtern auch in historischen Texten bietet Wordle. (Es ist hier allerdings zu beachten, dass zwar die vom Programm ermittelten Zahlenangaben stimmen, aber bei der graphischen Ausgabe dieser Werte – die keinen sprachwissenschaftlichen, sondern einen ästhetischen Zweck verfolgt – gerade die höchstfrequenten Wörter ausgeblendet werden.)

[33] Die dargestellte Methode wurde von Severin Stampler in seiner Diplomarbeit »Discrimination of Scribes in Medieval Manuscripts« (Graz 2009) unter anderem basierend auf Arbeiten von Bulacu und Schomaker sowie Srihari et al. entwickelt.

tursegmente einschließen, sowie die Winkel zweier Kontursegmentpaare, die einander horizontal oder vertikal unmittelbar gegenüberstehen.

Hinsichtlich der Strichstärke in bzw. Abständen zwischen Regionen werden sog. Lauflängen berechnet, das sind die Längen horizontaler bzw. vertikaler Segmente zwischen Konturpixeln entweder der Schrift selbst oder des Hintergrundes. Die Abstände innerhalb von Hintergrundregionen umfassen dabei die Ausmaße eingeschlossener Regionen sowie die Abstände zwischen Buchstaben, innerhalb von Vordergrundregionen werden Strichstärken und Buchstabenproportionen erfasst.

Alle diese Einzelmaße werden nun für die einzelnen Regionen eines Segmentes berechnet und in *normalisierten Histogrammen* je Segment zusammengefasst. Die normalisierten Histogramme aller Maße bilden in Summe einen sog. *Deskriptor*, der die charakteristischen Eigenschaften dieser Region beschreibt.

Dieser Deskriptor ist nun der Ausgangspunkt, um die eigentliche Detektion von Schreiberwechseln durchzuführen. Der Vorgang läuft wie folgt ab: Unter der Annahme, dass ein Schreiberwechsel am Beginn eines Abschnittes unwahrscheinlich ist, werden für eine vordefinierte Anzahl von teilweise überlappenden Segmenten am Beginn eines Abschnittes die einzelnen Deskriptoren berechnet und zu einem Gesamtmodell zusammengefasst. Dieses Gesamtmodell soll die charakteristischen Eigenschaften des präsumtiven Schreibers repräsentieren. Nun werden für die folgenden Segmente wiederum die Deskriptoren berechnet und über die *Ähnlichkeit* zwischen Gesamtmodell und Deskriptor für neue Segmente die Wahrscheinlichkeit der Zugehörigkeit berechnet. Ein Wechsel wird als erkannt angenommen, wenn genügend Deskriptoren vom aktuellen Modell stark abweichen, wobei der notwendige Grad der Abweichung ein Parameter der Methode ist.

Wurde ein Wechsel erkannt, so wird ein neues Gesamtmodell für den nächsten Schreiber aufgrund einer vorgegebenen Anzahl von Deskriptoren ab dem Schreiberwechsel erstellt und der Vorgang wiederholt sich. Wurde kein Wechsel erkannt, wird das aktuelle Gesamtmodell mit den für die neu hinzugekommenen Segmente berechneten Deskriptoren verfeinert. Alle diese Schritte laufen vollständig *automatisch* ab, also ohne die Notwendigkeit von Benutzereingaben.

Die Methode wurde mit unterschiedlicher Parametrisierung auf Datensätzen mit bis zu 16 verschiedenen (bekannten) Schreibern, getestet. Dabei konnten Erkennungsraten zwischen 80 und 91 % erzielt werden, wobei im besten Fall 92 % der vom System gemeldeten Wechsel korrekt waren.

Für die *Mikroebene* werden einzelne Buchstaben oder kurze Buchstabenfolgen (maximal im Ausmaß einzelner Wörter, nämlich der hochfrequenten Stempelwörter, vgl. Abschnitt 4.1) herangezogen und dafür folgende Deskriptoren berechnet:

- Abstand zwischen Basisgraph und diakritischem Zeichen
- Winkel zwischen Basisgraph und diakritischem Zeichen

Als Ausgangspunkt für die Bestimmung dieser beiden Deskriptoren wird der Punkt oben in der Mitte des Basisgraphs genommen und davon ausgehend zur Mitte der Region des diakritischen Zeichens gemessen.

- Strichstärke
- Eingeschlossene Fläche
  Diese beiden Deskriptoren werden mit dem oben beschriebenen Verfahren zur Berechnung von Lauflängen ermittelt. Für die Ermittlung des Maßes für die Strichstärke wird der Mittelwert eines Teiles der kurzen Lauflängen herangezogen.
- Schriftneigung
  Hierzu wird das Histogramm der Winkel zwischen Kontursegmenten und der Horizontalen in einer Region berechnet; der Schriftneigungsgrad ergibt sich als das Maximum im Histogramm dieser Werte.
- Visuelle Ähnlichkeit
  Für die Berechnung der visuellen Ähnlichkeit zwischen Buchstabenregionen gibt es eine Reihe von Methoden, basierend z. B. auf der Farbverteilung, den dominanten Farben oder der Textur des Bildes. Für das Projekt DA*mal*S hat sich die Verwendung so genannter *Kovarianz-Matrizen*, in denen statistische Informationen zur Farb- bzw. Grauwertverteilung innerhalb einer Region gesammelt werden, als am geeignetsten herauskristallisiert bzw. der Einsatz des SIFT-Deskriptors (Lowe), der die Häufigkeit von Richtungen in einer Region beschreibt.

Um von der Auflösung der Digitalisate unabhängig zu sein und somit codexübergreifende Analysen durchführen zu können, werden alle längen- bzw. flächenbezogenen Deskriptoren auf die Fläche der Region normalisiert.

Zur weiteren Analyse können entweder der Verlauf all dieser Deskriptoren einzeln oder die Unterschiede im Verlauf herangezogen werden. Zur Berechnung der Unterschiede wird im einfachen Fall von Einzelwerten (wie z. B. für Schriftneigung oder Winkel) der Absolutbetrag der Differenz dieser Einzelwerte, für komplexere Maße (wie z.B. beim visuellen Vergleich) speziell abgestimmte Unterschiedsmaße (deren Erklärung den Rahmen dieses Beitrages sprengen würde) herangezogen. Für die *Visualisierung* des Verlaufes können einfache Liniengraphiken verwendet werden. Im Projekt sind jedoch spezielle Methoden entwickelt worden, die den Verlauf eines Maßes in Form eines in der Transparenz modulierten Bandes darstellen, unter dem die Digitalisate durchscheinen und somit – wie schon am Ende von Abschnitt 4.1 dargestellt – der Verlauf gleichsam in die Bilder zurückprojiziert wird. Mit dieser Methode können auch mehrere Maße gleichzeitig dargestellt werden, wobei die Bänder auch überlappend angeordnet werden können und sich so der Effekt insgesamt verstärkt. Bei gleichzeitiger unterschiedlicher Färbung der Bänder bleibt die individuelle Interpretation der Einzelverläufe weiterhin möglich; siehe dazu Abb. 5 (Darstellung der Unterschiede zwischen den einzelnen <die>-Belegen anhand der Strichstärke des <d> im linearen Verlauf des Dokuments als transparenzmoduliertes Band über den Digitalisaten) und Abb. 6 (Darstellung der Unterschiede zwischen den einzelnen <die>-Belegen anhand der Strichstärke des <d>, der inneren Volumina des <d> sowie der Strichstärke des <i> als transparenzmodulierte Bänder über den Digitalisaten).

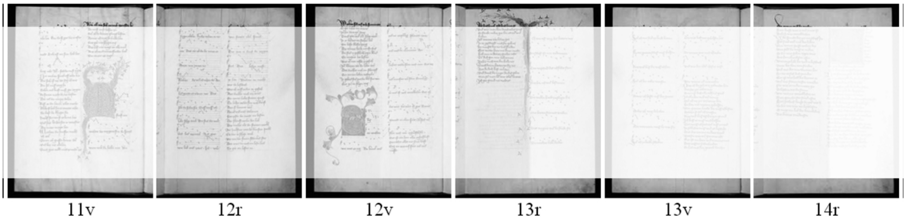


Abbildung 5. Darstellung eines Deskriptors im linearen Verlauf des Dokuments als transparenzmoduliertes Band über den Digitalisaten.

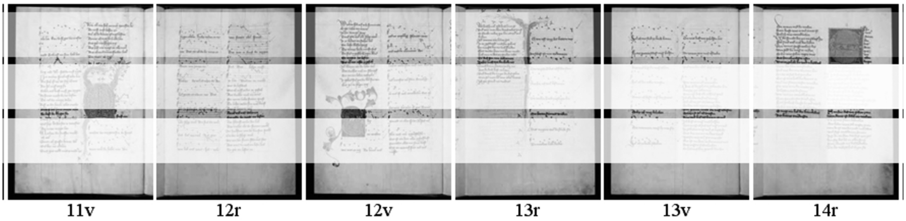


Abbildung 6. Darstellung dreier Deskriptoren im linearen Verlauf des Dokuments als transparenzmodulierte, teilweise überlappende Bänder über den Digitalisaten.

Für die komplexeren Maße oder Kombinationen von einfachen Maßen kann in einem weiteren Schritt unter der Annahme, dass die Maße aussagekräftig in Hinsicht auf die Schreiberhand sind, versucht werden, diese zu *gruppieren* (zu »clustern«). Hierzu werden die Distanzen zwischen den Maßen herangezogen, um Belege mit kleinen Distanzen zu Gruppen zusammenzufassen. Es können mehrere Gruppierungsvarianten mit einer unterschiedlichen Gruppenanzahl als Vorgabe berechnet werden. Stellt man nun die Zuordnung der einzelnen Vorkommen einer Buchstabenfolge über deren Position im Dokument als Verlauf dar, geben die wechselnden Zuordnungen über die Stabilität eines Clusters Auskunft. In Abb. 4 ist dies beispielhaft für das Wort »die« dargestellt.[34]

[34] Die Tabelle listet die ersten und letzten 10 Belege der Form <die> jedes der von Wackernell festgelegten Schreiberbereiche A, B, C, D. In der direkten Gegenüberstellung der Bildausschnitte treten mikroskopische Duktus-Unterschiede zwischen A und B deutlich zutage, während B und C in markanten Details (z. B.

Abb. 7 zeigt eine Anordnung der einzelnen <die>-Vorkommen basierend auf dem Unterschiedsmaß, Abb. 8 zeigt den Verlauf der Zuordnung zu Clustern bei der Vorgabe »drei Cluster« und Abb. 9 bei der Vorgabe »vier Cluster« zu bilden. Die Interpretation der Unterschiede in der Zuordnung zu Clustern zwischen Abb. 8 und Abb. 9 stärkt nun

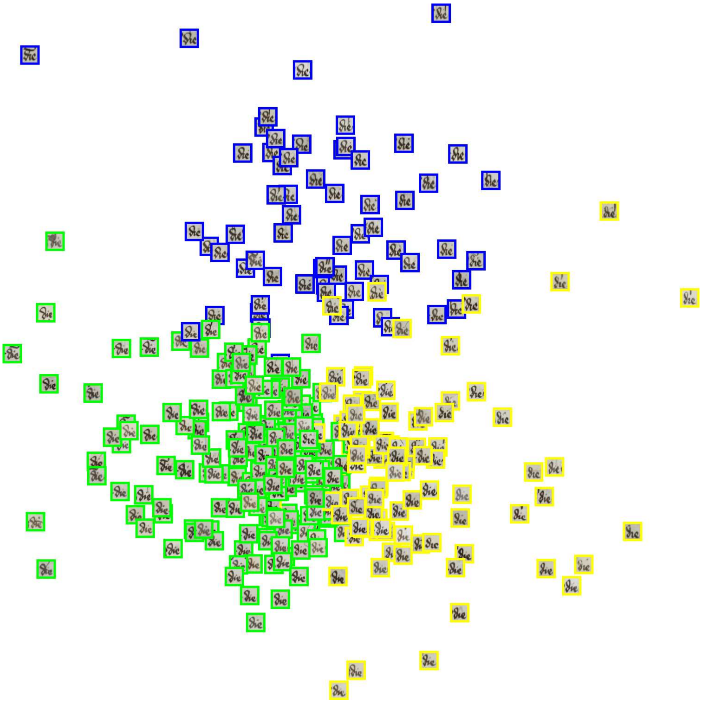


Abbildung 7. Kumulierende Zuordnung aller <die>-Belege im cpg 329 zu Clustern bei der Vorgabe »drei Cluster«.

Form und Position des i-Punktes) Übereinstimmung zeigen. Vgl. ausführlich Hofmeister und Hofmeister-Winter.

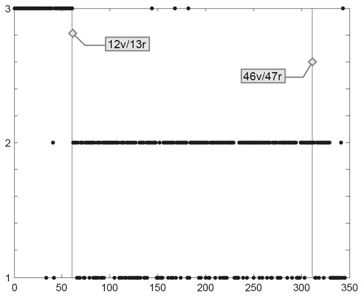


Abbildung 8. Lineare Zuordnung aller <die>-Belege im cpg 329 zu drei Clustern.

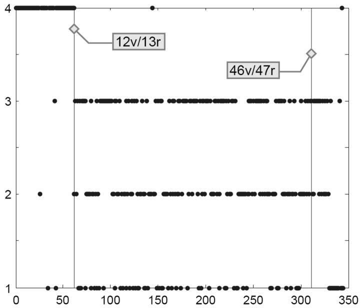


Abbildung 9. Lineare Zuordnung aller <die>-Belege im cpg 329 zu vier Clustern.

die schon von Wackernell gebildete Hypothese (vgl. Abschnitt 2.1), dass ein Schreiberhandwechsel in cpg 329 zwischen fol. 12v und 13r signifikant hervortritt, wogegen seine These eines weiteren Wechsels zwischen fol. 46v und 47r nicht erhärtet werden kann.

5 Ausblick: Weiterentwicklung von DA*mal*S mit Integrierung in MOSES

Das methodisch anpassungsfähige Projektdesign von DA*mal*S – rund um seine (bislang einzigartige) *mikrographetische* Bild-Text-Erfassung – sollte es erlauben, zahlreiche weitere Schriftdokumente auf allfällige Schreiberhandwechsel hin zu untersuchen, und zwar prinzipiell auch für andere Sprachen. Damit könnte eine Art *Fahndungskartei* heranwachsen, welche wiederum die Grundlage für die Zeichnung einer historischen ›Wanderkarte‹ von Schriftdokumenten quer durch Europa bilden mag.

Das *Limit* für die Authentifizierung von Schreiberhänden wird von DA*mal*S erst dort erreicht, wo sich in den Schriftzügen (z. B. wegen dominanter kalligraphischer Bestrebungen oder materieller Einschränkungen) keine ausreichenden Spuren von psychomotorischen Einprägungen nachweisen lassen. Aus dieser Einsicht kann per Umkehrschluss und unter Zugrundelegung einer diachronen Betrachtungsweise für eine Schriftkultur zugleich der Beginn ihrer *Individualisierung* eruiert werden, also der Zeitpunkt, ab wann innerhalb einer Sprachgemeinschaft deren Verschriftungstechnik so etwas wie einen schreiberspezifischen ›Fingerabdruck‹ ausgebildet hatte.

Bestärkt durch das große (sogar europaweite) Medien-Echo, das die soeben geschilderten Entwicklungsperspektiven von DA*mal*S mittlerweile erfahren haben,[35] wurde

[35] Vgl. die auswahlhafte Dokumentation der Print- und Filmberichte über DA*mal*S seit einer Presseaussendung im März 2008 auf der Projekt-Homepage.

vom Projektleiter das ›Meta-Projekt‹ MOSES (›Musterorientiertes System zur Erfassung von Schriftindividualität‹) konzipiert. Es soll DA*mal*S integrieren, aber darüber hinaus auch für *gegenwartsbezogene* Problemlösungen im Zusammenhang mit der Verifizierung von individuellen Schriftzügen offen sein. Durch entsprechende Verfeinerungen und sprachspezifische Anpassungen der Authentifizierungstechnik mag es möglich werden, z. B. auch der *forensischen Schriftforschung* zu dienen; praxisnahe Vorstudien finden dazu bereits statt.[36]

Eine erfreulich konkrete Aussicht auf die Umsetzung von MOSES hat sich Anfang 2009 durch seine Einbindung in eine Initiative der Ludwig Boltzmann-Gesellschaft ergeben. Diese strebt die Entwicklung einer (weltweit vernetzten) ›Archiv-, Text- und Editionstheorie‹ an. Sollte dieser Antrag – trotz einer zur Zeit (global wie national) sehr angespannten Finanz- und Förderungslage – erfolgreich sein, ließe sich 2010 in Graz ein eigenes MOSES-Forschungsinstitut gründen; durch offizielle ›Letters of Intent‹ an die LBG sowohl von der Karl-Franzens-Universität Graz als auch von der Forschungsgesellschaft Joanneum Research wurden dafür jedenfalls die ersten Schritte gesetzt.

Bibliographie

Aitchison, Jean. *Wörter im Kopf. Eine Einführung in das mentale Lexikon.* Aus dem Englischen von Martina Wiese. Tübingen: Niemeyer, 1997.

Arbeitsgemeinschaft für germanistische Edition. <http://www.ag-edition.org/html/archiv.html>.

Bibliotheca Palatina – digital. UB Heidelberg. <http://www.ub.uni-heidelberg.de/helios/digi/palatina-digital.html>.

Bromm, Gudrun. »Neue Vorschläge zur paläographischen Schriftbeschreibung.« *Methoden der Schriftbeschreibung.* Hrsg. Peter Rück. Stuttgart: Thorbecke, 1999. 21–43.

Bulacu, Marius and Lambert Schomaker. »Text-independent writer identification and verification using textural and allographic features.« *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.4 (2007): 701–17.

Cappelli, Adriano. *Lexicon abbreviaturarum.* 2., verb. Aufl. (Nachdr. d. Ausg. Leipzig 1928.) München: Hieronimus, 1988.

[*cpg 329, Online-Präsentation*]. <http://diglit.ub.uni-heidelberg.de/diglit/cpg329>.

DAmalS, Projekthomepage. <http://www.uni-graz.at/wernfried.hofmeister/damals>.

Das Brixner Dommesnerbuch. Mit elektronischer Rohtextversion und digitalem Voll-

[36] Das gesamte Autorenteam dieses Beitrags ist eingeladen, im Juni 2009 auf dem Kriminologenkongress der Gesellschaft für Forensische Schriftuntersuchung (siehe Tagungsprogramm) die bislang entwickelte Leistungsfähigkeit von DA*mal*S anhand der Befundung strittiger handschriftlicher Dokumente unter Beweis zu stellen, und hofft seinerseits, dabei neue Entwicklungsimpulse zu erhalten.

faksimile auf CD-ROM. Hrsg. Andrea Hofmeister-Winter. Innsbruck: Institut für Germanistik, 2001.

Gesellschaft für Forensische Schriftuntersuchung. <http://www.gfs2000.de>; [*Informationen zum Tagungsprogramm*]. <http://www.gfs2000.de/main_d1.htm>.

Grabowski, Joachim. »Bedingungen und Prozesse der schriftlichen Sprachproduktion.« *Psycholinguistik/Psycholinguistics. Ein internationales Handbuch.* Hrsg. Gert Rickheit, Theo Herrmann und Werner Deutsch. Berlin: de Gruyter 2003. 355–368.

Grun, Paul A. *Schlüssel zu alten und neuen Abkürzungen.* (Reprint d. Ausg. 1966.) Limburg/Lahn: Starke, 2002.

Hofmeister, Wernfried. »Die Edition als ›offenes Buch‹: Chancen und Risiken einer Transponierungs-Synopse, exemplarisch dargestellt an der Dichtung *Von des todes gehugede* des sog. Heinrich von Melk.« *Produktion und Kontext. Beiträge der Internationalen Fachtagung der Arbeitsgemeinschaft für germanistische Edition im Constantijn Huygens Instituut, Den Haag, 4. bis 7. März 1998.* Hrsg. H. T. M. van Vliet. Tübingen: Niemeyer 1999. 23–39.

Hofmeister, Wernfried. »Der Mut zur Lücke: Auf den Spuren von Textnachträgen in der Manessischen Liederhandschrift. Ein Beitrag zu einer ›Überlieferungs-Philologie‹ des Mittelalters.« *Entstehung und Typen mittelalterlicher Lyrikhandschriften. Akten des Grazer Symposiums, 13.–17. Oktober 1999.* Hrsg. Anton Schwob und András Viskelety unter Mitarbeit von Andrea Hofmeister-Winter. Bern: Lang, 2001. 79–106.

Hofmeister, Wernfried. »Ein Autor ›outet‹ sich: Hugo von Montfort (1357–1423) im rezeptionellen Spannungsfeld.« *Autor – Autorisation – Authentizität. Beiträge der internationalen Fachtagung der Arbeitsgemeinschaft für Germanistische Edition in Verbindung mit der Arbeitsgemeinschaft Philosophischer Editionen und der Fachgruppe Freie Forschungsinstitute in der Gesellschaft für Musikforschung, Aachen, 20. bis 23. Februar 2002.* Hrsg. Thomas Bein, Rüdiger Nutt-Kofoth und Bodo Plachta. Tübingen: Niemeyer, 2004. 165–72.

Hofmeister, Wernfried. »Perspektiven und Auswirkungen des Edierens am Beispiel der neuen Hugo von Montfort-Ausgabe.« *Aller weishait anevang / ist ze brûfen an dem ausgang. Akten des Symposiums zum 650. Geburtstag Hugos von Montfort, Dornbirn, 19.–22. September 2007.* Hrsg. Klaus Amann und Elisabeth De Felip-Jaud (im Druck).

Hofmeister, Wernfried und Andrea Hofmeister-Winter. »Schriftzüge unter der High-Tech-Lupe. Theoretische Grundlagen und erste praktische Ergebnisse des Grazer Forschungsprojekts DAmalS (›Datenbank zur Authentifizierung mittelalterlicher Schreiberhände‹).« *editio. Internationales Jahrbuch für Editionswissenschaft* 22 (2008): 90–117.

Hofmeister-Winter, Andrea. *Das Konzept einer »Dynamischen Edition« dargestellt an* *der Erstausgabe des »Brixner Dommesnerbuches« von Veit Feichter (Mitte 16. Jh.). Theorie und praktische Umsetzung.* Göppingen: Kümmerle, 2003.

Hofmeister-Winter, Andrea. »Textkritik als Erkenntnisprozeß: sehen – verstehen – deuten.« *editio. Internationales Jahrbuch für Editionswissenschaft* 19 (2005): 1–9.

Hofmeister-Winter, Andrea. »Die Grammatik der Schreiberhände. Versuch einer Klärung der Schreiberfrage anhand der mehrstufig-dynamischen Neuausgabe der Werke Hugos von Montfort.« *Edition und Sprachgeschichte. Baseler Fachtagung 2.–4. März 2005.* Hrsg. Michael Stolz in Verbindung mit Robert Schöller und Gabriel Viehhauser. Tübingen: Niemeyer, 2007. 89–116.

Hofmeister-Winter, Andrea. [*Forschungshomepage*]. <http://www.uni-graz.at/~hofmeisa>.

Hugo von Montfort. Mit Abhandlungen zur Geschichte der deutschen Literatur, Sprache und Metrik im XIV. und XV. Jahrhundert. Hrsg. J[osef] E[duard] Wackernell. Innsbruck: Wagner, 1881.

Hugo von Montfort. Das poetische Werk. Hrsg. Wernfried Hofmeister. Mit einem Melodie-Anhang von Agnes Grond. Berlin: de Gruyter, 2005.

Hugo von Montfort, [Editions-Homepage]. <http://www-gewi.uni-graz.at/montfort-edition>.

Institut für Informatik/Abteilung für Sprachverarbeitung an der Universität Leipzig. [*Projekt »Wortschatz«*]. <http://wortschatz.uni-leipzig.de>.

Krämer, Sigrid. *Scriptores codicum medii aevi. Datenbank von Schreibern mittelalterlicher Handschriften.* CD-Rom und Beiheft. Augsburg: Rauner, 2003.

Lowe, David G. »Distinctive image features from scale-invariant keypoints.« *International Journal of Computer Vision* 60.2 (2004): 91–110.

Parzival-Projekt. <http://www.parzival.unibe.ch>.

Produktion und Kontext. Beiträge der Internationalen Fachtagung der Arbeitsgemeinschaft für Germanistische Edition im Constantijn Huygens Instituut, Den Haag, 4. bis 7. März 1998. Hrsg. H. T. M. van Vliet. Tübingen: Niemeyer, 1999.

Schieb, Gabriele. »Editionsprobleme altdeutscher Texte.« *PBB* [Halle/S.] 89 (1967): 404–30.

Schlögl, Waldemar. Rez. von *L'expertise des écritures médiévales. Recherche d'une méthode avec application à un manuscrit du XIe siècle: Le lectionnaire de Lobbes, Codex Bruxellensis 18018, Gand 1973 [...], von Léon Gilissen. Deutsches Archiv für Erforschung des Mittelalters* 33 (1977): 264–65.

Schneider, Karin. *Paläographie und Handschriftenkunde für Germanisten. Eine Einführung.* Tübingen: Niemeyer, 1999.

Schneider, Karin. »Akzentuierung in mittelalterlichen deutschsprachigen Handschriften.« *Edition und Sprachgeschichte. Baseler Fachtagung 2.–4. März 2005.* Hrsg. Michael Stolz in Verbindung mit Robert Schöller und Gabriel Viehhauser. Tübingen: Niemeyer, 2007. 17–24.

Schubert, Martin. »Sprechende Leere. Lücke, Loch und Freiraum in der Großen Heidelberger Liederhandschrift.« *editio. Internationales Jahrbuch für Editionswissenschaft* 22 (2008): 118–38.

Spechtler, Franz V. *Die Heidelberger Handschrift cpg 329 und die gesamte Streuüberlieferung.* In Abbildung hrsg. von Eugen Thurnher, Franz V. Spechtler und Ulrich Müller. Göppingen: Kümmerle, 1978. 12–20.

Srihari, Sargur N., Sung-Hyuk Cha, Hina Arora, and Sangjik Lee. *Handwriting identification: Research to study validity of individuality of handwriting and develop computer-assisted procedures for comparing handwriting.* Buffalo (NY): University at Buffalo, 2001.

Wege zum Text. Überlegungen zur Verfügbarkeit mediävistischer Editionen im 21. Jahrhundert, Grazer Kolloquium, 17.-19. September 2008. Hrsg. Wernfried Hofmeister und Andrea Hofmeister-Winter. Tübingen: Niemeyer, 2009 (im Druck).

»Wege zum Text«, [Tagungshomepage]. <http://www.uni-graz.at/wernfried.hofmeister/wegezumtext>.

Welker, Lorenz. »Die Melodien des Burkhard Mangold.« *Hugo von Montfort. Einführung zum Faksimile des Codex Palatinus Germanicus der Universitätsbibliothek Heidelberg.* Mit Beiträgen von Franz Viktor Spechtler u. a. Wiesbaden: Reichert, 1988. 47–60.

Werner, Wilfried. »Die Handschrift und ihre Geschichte.« *Hugo von Montfort. Einführung zum Faksimile des Codex Palatinus Germanicus der Universitätsbibliothek Heidelberg.* Mit Beiträgen von Franz Viktor Spechtler u. a. Wiesbaden: Reichert, 1988. 7–11.

Wordle (Programm). <http://www.wordle.net>.

Wortfrequenzlisten der Universität Leipzig. <http://wortschatz.uni-leipzig.de/Papers/top10000de.txt>.

Zentrum für Informationsmodellierung in den Geisteswissenschaften (ZIMig) an der Karl-Franzens-Universität Graz. <http://www.uni-graz.at/zim>.

Digital Palaeography*

Mark Aussems, Axel Brink

Abstract

This article seeks to explore new digital ways of distinguishing between scribal hands in medieval manuscripts. An analysis of traditional palaeographical approaches to hand identification will be followed by a discussion in which attention will be paid both to the use of computer software to enhance existing methods of scribal identification, and to the benefits of *Quill*, an innovative automatic writer identification tool. A case study involving a manuscript of the collected works of Christine de Pizan (London, British Library, Harley 4431) will serve to demonstrate that traditional palaeographical methods of analysing scribal hands can greatly benefit from the use of specialised computer software.

Zusammenfassung

Der Beitrag versucht, neue digitale Wege zu erkunden, um Schreiberhände in mittelalterlichen Handschriften zu unterscheiden. Nach einer Analyse des traditionellen paläographischen Ansatzes zur Handidentifikation diskutiert er sowohl die Möglichkeiten, bisherige Methoden mit dem Computer zu verbessern als auch den Nutzen von *Quill*, eines innovativen Werkzeugs zur automatischen Schreiberidentifikation. Eine Fallstudie einer Handschrift der gesammelten Werke der Christine de Pizan (London, British Library, Harley 4431) demonstriert, dass traditionelle paläographische Methoden der Schreibhandanalyse von der Anwendung spezialisierter Computersoftware deutlich profitieren können.

1 Introduction

The identification of scribal hands in medieval manuscripts is one of the most important problems in the discipline of palaeography. Over the years, experienced palaeographers have created many methodological approaches to scribal identification, each with its own advantages and inconveniences. These methodologies are without exception based

* The authors would like to thank Prof James C. Laidlaw (University of Edinburgh) for his useful comments on an earlier draft of this article, and the British Library for their kind permission to use images of MS Harley 4431.

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 293–308.

on what could be called traditional palaeography. They promote the use of traditional palaeographical tools such as protractors and set squares to measure letter heights and widths, distances between characters, margins, and angles of inclination. Furthermore, they consider certain immeasurable features of a medieval hand—such as *ductus*, writing speed, and the number and types of abbreviations that are used—as criteria that can be used to distinguish between scribal hands.

In recent years, a significant number of digitisation projects involving medieval manuscripts have seen the light of day. This ever-increasing corpus of digital manuscript images has made palaeographers, codicologists, philologists, and art historians around the globe aware of the advantages which it may generate for the study of manuscripts. The availability of high resolution manuscript images has also inspired computer scientists to contribute to the palaeographical discussions about hand identification. This cross-fertilisation of information technology and palaeography, albeit in a preliminary stage, has produced some interesting outcomes, which have not yet received the attention they deserve.

The present article seeks to analyse several of the aforementioned traditional methods of scribal identification in the light of the new digital possibilities within the field of manuscript analysis. To what extent will traditional palaeography be able to benefit from recently developed computer software? Can palaeographical measurements be carried out faster and more accurately with the arrival of the computer on the palaeographer's desk? Are there new possibilities for manuscript research to be discovered that had not been possible before?

Our attention will centre on *Quill*, a promising outcome of the recent involvement of computer experts within the field of palaeography. An analysis of the scribal differences in a manuscript containing works by the French author Christine de Pizan (*ca* 1364 – *ca* 1430)—whose manuscripts are subject to heated discussions between specialists about the number of hands that can be distinguished in them—will be carried out conjointly by *Quill* and through a traditional palaeographical method. The results of this case study can provide useful insights into the benefit of automated and digital recognition of scribal hands to the discipline of palaeography.

2 Traditional Palaeography

When, around the turn of the seventeenth century, the Benedictine monks of the Congregation of St. Maur started cataloguing specimens of handwritten texts, they little knew that some three hundred years later, their approach to manuscript texts would become one of the most important auxiliary sciences within the discipline of History. Although he did not coin the term 'palaeography', Jean Mabillon was the first scholar to create a set of chronologically ordered samples of handwriting (Mabillon). Mabillon's fellow-brothers Bernard de Montfaucon, Charles-François Toustain, and René-Prosper Tassin continued and expanded his work, thus creating the first systematic survey of handwriting (Montfaucon; Toustain and Tassin). From the second half of the nineteenth century onwards, this approach to medieval manuscripts became popular and was studied extensively by, among others, Léopold Delisle, Ludwig Traube, Wilhelm Wattenbach, and, in more recent times, Jean Mallon, Bernhard Bischoff, Léon Gilissen, and Albert Derolez.[1]

Within this palaeographical framework, more and more attention was paid to the history—or the *archaeology*—of the medieval book. Under the influence of Alphonse Dain, Charles Samaran, François Masai, and Léon Delaissé, the discipline of codicology was created, its objective being to study the history and the production of the medieval codex (Dain; Delaissé). This interest in the medieval book as a material and cultural object brought about a change in palaeography. Influenced by codicological studies on the functioning of medieval scriptoria, palaeographers became more and more interested in distinguishing between medieval hands in a single text or a corpus of texts rather than between different scripts. This was the beginning of a structured and logical analysis of medieval handwriting as a phenomenon, influenced not only by external factors such as the education of the scribes, the size and material of the quill, and the writing support, but also by the scribes' own subconscious execution of a particular script.

Over the last fifty years, a number of methods have been created that can be used to distinguish between multiple hands that make use of the same script to transcribe a text or a corpus of texts. In 1952, Jean Mallon, in his work *Paléographie romaine*, introduced a list of seven aspects of a medieval hand that should be taken into account when distinguishing scribal hands (Mallon 23):[2]

- the *form*, the morphology of the letters;
- the *angle of writing* in relation to the line for writing;
- the *ductus*;
- the *modulus*, the dimensions of the letters;
- the *contrast*, the difference in thickness between hair lines and shadow lines;
- the *writing support*;
- the *internal characteristics*, the nature of the text.

This list of what could be called *differentiators* has been used by other palaeographers and has become the basis of later methodologies. Léon Gilissen adapted and developed Mallon's differentiators in his 1973 work *L'expertise des écritures médiévales.* Gilissen drops the last two differentiators in Mallon's list in favour of another, called *style*, by which he means "une manière d'être qui se répercute sur tous les éléments de l'écriture,

[1] See the Bibliography at the end of this article for references to works of these scholars.

[2] See also Stokes in this volume.

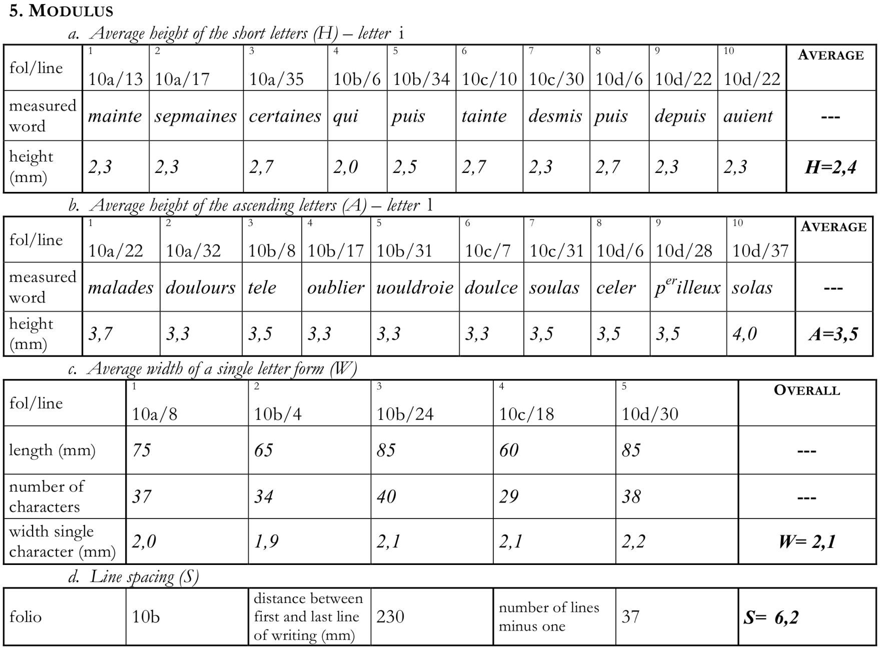
qui affecte et qui marque le phénomène entier" [a feature that has repercussions on all aspects of writing, thus affecting and marking the entire phenomenon] (50). Furthermore, Gilissen tries to make the methodology more objective and more accurate by expressing some of the differentiators in numerical values rather than by giving lengthy descriptions. In 1995, the Dutch palaeographer Jan Burgers surveyed the existing methodologies for hand identification, including some interesting publications in the field of forensic analysis (e.g. Michel; Hardy and Fagel). He combined these approaches in what could be called the *Burgers methodology*, a successful method of differentiating between scribal hands in charters (Burgers). In recent years, Mark Aussems has shown that this method can—with some adjustments—be applied to scribal hands in medieval manuscripts (Aussems 2006; Aussems 2008). He also coined the term *scribal fingerprint* to denote a set of objective, accurate, and quantifiable characteristics that are unique to one particular scribal hand (Aussems 2006 10). To this extent, he emphasises the benefit of computers and specific computer software to the quantitative study of medieval handwriting (Aussems 2007; Aussems 2008).

3 Digital Palaeography

The expression *digital palaeography* was coined by Arianna Ciula in her 2005 article 'Digital palaeography: using the digital representation of medieval script to support palaeographic analysis' and was created as a result of cross-fertilisation between the academic disciplines of Palaeography, Computing, and Artificial Intelligence. The many digitisation projects involving medieval manuscripts have undoubtedly contributed to this collaboration. By way of a definition, digital palaeography is the discipline that makes use of computers and computer software to analyse classical and medieval handwriting. It thereby relies on the quantitative aspect of palaeography; the values attached to the aforementioned differentiators need to be turned into numerical data. Of the list provided in Aussems 2006, four differentiators can be expressed as numerical values: the *angle of inclination*, the *angle of writing*, the *modulus*, and the *degree and type of cursivation of connecting characters*.

The advantages of numerical data lie in the fact that they enable rapid analysis by computer software, and enable—and indeed even facilitate—a comparison between different hands that is based on objective and comparable data rather than on 'verbal' descriptions of a hand. These advantages become clear when we compare the two examples below: a verbal description of X, one of the hands found in the original manuscripts of works of Christine de Pizan and, in Figure 1, part of a numerical description of one of the hands in Christine's *Queen's Manuscript* (London, British Library, Harley 4431).

> Mieux vaut caractériser cette main par des tendances que par des formes : en effet,



5. MODULUS

a. Average height of the short letters (H) – letter i

fol/line	1 10a/13	2 10a/17	3 10a/35	4 10b/6	5 10b/34	6 10c/10	7 10c/30	8 10d/6	9 10d/22	10 10d/22	AVERAGE
measured word	*mainte*	*sepmaines*	*certaines*	*qui*	*puis*	*tainte*	*desmis*	*puis*	*depuis*	*auient*	---
height (mm)	*2,3*	*2,3*	*2,7*	*2,0*	*2,5*	*2,7*	*2,3*	*2,7*	*2,3*	*2,3*	***H=2,4***

b. Average height of the ascending letters (A) – letter l

fol/line	1 10a/22	2 10a/32	3 10b/8	4 10b/17	5 10b/31	6 10c/7	7 10c/31	8 10d/6	9 10d/28	10 10d/37	AVERAGE
measured word	*malades*	*doulours*	*tele*	*oublier*	*uouldroie*	*doulce*	*soulas*	*celer*	*p^{er}illeux*	*solas*	---
height (mm)	*3,7*	*3,3*	*3,5*	*3,3*	*3,3*	*3,3*	*3,5*	*3,5*	*3,5*	*4,0*	***A=3,5***

c. Average width of a single letter form (W)

fol/line	1 10a/8	2 10b/4	3 10b/24	4 10c/18	5 10d/30	OVERALL
length (mm)	*75*	*65*	*85*	*60*	*85*	---
number of characters	*37*	*34*	*40*	*29*	*38*	---
width single character (mm)	*2,0*	*1,9*	*2,1*	*2,1*	*2,2*	***W= 2,1***

d. Line spacing (S)

folio	10b	distance between first and last line of writing (mm)	230	number of lines minus one	37	***S= 6,2***

Figure 1. A (slightly altered) numerical description of the *modulus* of hand A in manuscript London, British Library, Harley 4431; Aussems 2006, Appendix H, p. 6.

> beaucoup plus habile que les deux autres, elle est, par voie de conséquence, également plus changeante, adoptant des styles d'écriture assez variés qui relèvent tous, néanmoins, de la cursive livresque. La constante la plus remarquable de X, c'est une exubérance, un goût pour les grandes envolées de la plume se déployant en un foisonnement de boucles et de volutes et faisant alterner pleins et déliés. (Ouy and Reno 1980 226)

Another advantage of the use of computer software in the field of quantitative palaeography is the increase in accuracy of the measurements that are carried out within a manuscript. Whereas traditional palaeographers had to revert to the use of a ruler or a set square to measure letters on the pages of a manuscript—thereby obviously having to operate very circumspectly so as not to damage the codex—this rather painstaking task can nowadays be executed quickly, safely and accurately by means of special imaging software that has a built-in measuring tool which allows palaeographers to carry out digital measurements.[3]

So far, we have focused on the added value of computers for the field of palaeography in its traditional form. However, in recent years there have been a number of very interesting developments within the disciplines of Artificial Intelligence and Computing which have the capacity not only to contribute to current palaeographic methodologies, but also to enhance or renew them, while *en passant* changing the way we look at scribal hands and medieval handwriting. One of the most promising initiatives in this context is *Quill*, a program to automatically compute writer features in historical documents.

4 *Quill*: A Different Approach to Digital Palaeography

Researchers at the Artificial Intelligence department at the University of Groningen have shown that one of their established automatic techniques can be used to identify medieval scribes (Bulacu and Schomaker 2007). Now, as part of the NWO (The Netherlands Organisation for Scientific Research) project TRIGRAPH,[4] a new automatic technique is emerging that focuses on historical handwriting in particular: the *Quill Dynamics Feature* (or: *Quill*).[5] Although the paper describing this technique in detail is expected to appear only in 2010, *Quill* is already contributing to several historical research initiatives.

Quill relies on the principle that writing instruments with an oblong contact surface (such as a quill) introduce writer-specific variation in the width of the ink trace. In a scanned document, *Quill* measures “the relation between the local direction and width of the ink traces” (Brink et al. 2010, section I). This is done by first determining the contours of the text: trajectories of dark (ink) pixels adjacent to light (parchment) pixels. Then, the contours are traversed counter-clockwise, while performing two measurements at every pixel: the direction of the ink and its width. This results in thousands of direction-width measurements for a document. The measurements are agglomerated in a *Quill Probability Distribution (QPD)*, which expresses the frequency at which each ink width was produced at each ink direction. Such a QPD is specific for each hand.

The intended use of *Quill* is to be incorporated in a system that performs writer identification. Given a query document, such a system can search a corpus of handwritten

[3] Computer programs like *Adobe PhotoShop* and *The GIMP* (short for: *GNU Image Manipulation Program*) already have built-in tools to measure distances and angles on-screen.

[4] The TRIGRAPH project (2005–2009) aims to improve automatic writer identification methods for forensic application by combining manual, semi-automatic, and fully-automatic methods. Recent developments have broadened the project to include historical handwriting as well.

[5] See Brink et al. 2010. Since this article is forthcoming, references will be given to its numbered sections rather than to page numbers.

documents by comparing the handwriting of the query document to that of all the other documents. It then returns a sorted list of documents with similar handwriting. The comparisons are based on an automatically computed feature such as a QPD, computed by *Quill*. In this way, given a document, a *hit list* is suggested of other documents that could have been written by the same hand. Although never 100% correct, this method works very well and saves the palaeographer a lot of time.

Writer identification by computing a hit list is not the only possibility. Another option is to apply a *clustering technique* to distinguish clusters (groups) of documents with similar handwriting, based on the QPD. This is very interesting when working with a corpus of texts, where it is known how many scribes were involved in their transcription—the clustering technique can be used to order the documents as if they were written by *k* scribes, where *k* is a number chosen by the user. An important issue is that if the handwriting in the analysed documents is quite similar, clustering will yield different groups in repeated executions. The obvious reason is that a strict separation is hard to make in such a case. A possible solution is to repeat the clustering many times while keeping track of the number of times every pair of documents were clustered into the same group.

5 Case Study

By means of an experiment, *Quill* was tested on a set of images of Dutch charters (1299-1328) in an earlier stage and proved to yield results that equal or surpass those of comparable computer programs (Brink et al. 2010, sections V and VI).[6] In the present article, we will test *Quill* on a different set of documents, but also examine whether the results obtained by *Quill* match those yielded by a palaeographical analysis of the same set of documents. Subject of this case study are two parts of the famous *Queen's Manuscript* (London, British Library, Harley 4431), a collection of thirty works by French author Christine de Pizan produced in the closing months of 1413 and presented to Queen Isabeau of France in January 1414 (Laidlaw 2005). The manuscript is unique in that for many of Christine's texts—which she often revised and corrected as they went through multiple 'editions'—Harley 4431 constitutes the last known version. Moreover, research conducted on the number of scribes that were involved in the transcription process of the codex has yielded divergent conclusions: whereas Gilbert Ouy and Christine Reno conclude that the Queen's Manuscript has been transcribed in its entirety by a single scribe (Ouy and Reno 1980), Sandra Hindman, James Laidlaw and others have taken a different stand (Hindman 1983, Laidlaw 1983 and 1987, Aussems 2006).

[6] Comparable applications include *Hinge* (Bulacu and Schomaker, Text-independent Writer Identification) and *Fraglets* (Schomaker et al. 2004).

Of particular interest in this case is quire 6 of the Harley MS, which is discussed in detail in Laidlaw 1987 and Aussems 2006. This quire contains folios 44 to 50, 50bis, 50ter, 51 and 52. It was originally conceived as a regular quaternion (44-47 / 48, 50ter, 51-52), but enlarged by a binion of which the third leaf was subsequently cancelled (49-50 / *, 50bis).[7] What is more, folios 50bis and 50ter are left blank. Figure 2 is a graphical representation of the structure of this quire. However, it is not just the codicological composition of quire 6 that interests us. In terms of scribal hands, the folios of this quire are witnesses to what seems to be a very intriguing change of hands. The first two folios of this quire—44, 45, and 46a—have been transcribed by the hand we have come to call B. Column 46b, however, seems to have been written not only in a paler ink, but also by a different hand, which we will call A. The verso of folio 46 is, again, written by hand B, who continues up to and including folio 48a. Folios 48b up to and including 51a, then, seem to be copied in hand A, after which B probably takes over again and finishes the quire. Figure 5 presents this change of hands by means of a diagram, and Figures 3 and 4 show folios 48r and 51r respectively, which clearly demonstrate the alleged change of hands from B to A and vice versa.

By means of a case study, we will subject these folios to a thorough palaeographical examination, carried out both by *Quill* and by using a computer-enhanced traditional approach. To this extent, we divided the folios into sections A and B, each corresponding to their respective assumed scribe. Consequently, we took samples of each hand by selecting parts of the text that do not contain miniatures, illumination or decoration.[8] The final result of this operation is seven samples of hand A and seven samples of hand B, each amounting to 170 lines of text (see Figure 6).[9] All text specimens were taken from columns b and c of the aforementioned folios only, because the text in columns a and d always appears slightly curved and oblique in the photos due to the impossibility of fully opening the large codex. Thus, we eliminate the possibility that the distorted text in columns a and d might corrupt our test results.

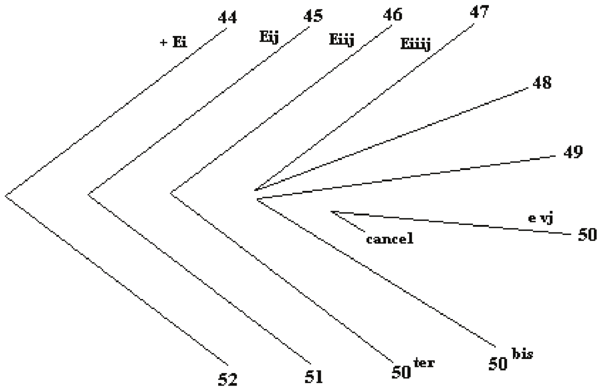
5.1 Traditional Palaeography

The 'traditional' palaeographical analysis involved an examination of the angle of inclination, and the height of the ascending characters of the handwriting in each of the samples. It was carried out by using the computer program *GIMP* to measure distances and angles according to the methodology as set out in Aussems 2006. The angle of

[7] In the two schematic representations of the quires, the sign / indicates the heart of the quire; * means a cancelled folio.

[8] In order for *Quill* to be as accurate as possible, every element that does not constitute a part of the text needs to be removed from the text that is to serve as input.

[9] The samples were taken from the following folios. Hand A: 46b:25 (lines), 48b:31, 48c:15, 49b:36, 49c:21, 50b:19, 50c:23. Hand B: 44b:27, 45b:10, 46c:22, 47b:28, 51c:30, 52b:39, 52c:14.



a. 'V' (i.e. 5) is written at the foot of fol. 48v, close to the binding, and 'No^{u}' is written in a fifteenth-century hand in the equivalent position on fol. 49r. Both these marks were probably instructions to the binder.
b. Bindings strings are visible between fols 47 and 48, and between fol. 50 and the cancel.

Figure 2. Graphical representation of the codicological structure of quire 6. Source: Aussems 2006 85; Laidlaw 1987 63.

inclination was measured at the right hand side of ten different shafts of the character *l*. The modulus consists of four different elements, one of which is the height of the ascending characters: it was determined by measuring ten characters *l*.[10] Finally, the measurements of each hand were averaged. The results are presented in Figure 7.

What becomes clear—not only from the averages presented above, but also from the results of each separate sample—is that there seems to be a substantial difference between hand A and hand B. Whereas the ascenders in the handwriting of scribe A are almost at right angles to the base line, those in hand B seem to be on average 10° out of the vertical. Furthermore, these ascenders are generally 0.5 mm larger in hand B. This is quite a significant difference, given the small standard deviations and the fact that the average numbers were calculated on the basis of 70 measurements for each hand.

[10] An inconvenience related to the images of Harley 4431, taken by the British Library photographic staff, is that the distance between the camera and the manuscript was not fixed, meaning that the readings—be they in millimetres or in pixels—cannot be compared. In order to overcome this problem, we measured the distances in pixels and subsequently converted them to millimetres by using the ruler placed within each photograph, giving us the exact distance in millimetres.

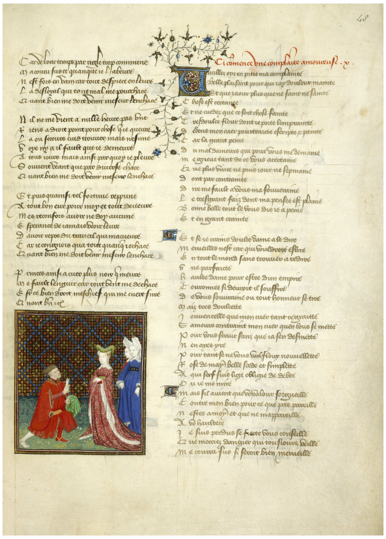


Figure 3. Folio 48r of the Queen's MS, showing hands B (column a) and A (column b). © British Library Board. All Rights Reserved. MS Harley 4431.

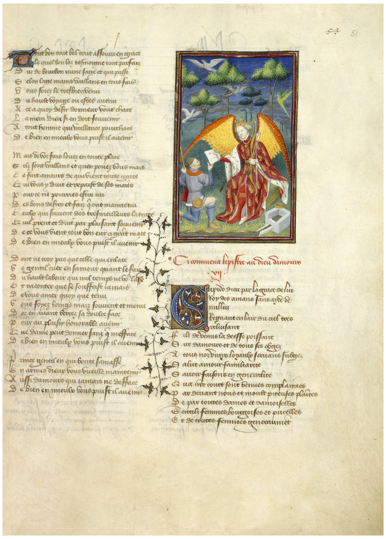


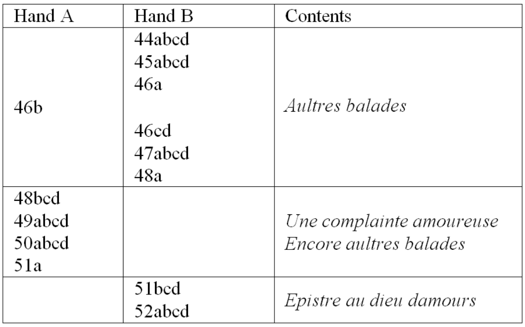
Figure 4. Folio 51r of the Queen's MS, showing hands A (column a) and B (column b). © British Library Board. All Rights Reserved. MS Harley 4431.

5.2 Quill

The text fragments were also subjected to a clustering experiment. To be specific, the clustering technique used is called *k-means*. As anticipated, the handwriting in the documents is too similar for a reliable instant grouping, thus the clustering was repeated 10,000 times. Figure 8 shows how often each pair of documents was classified in the same group, on a scale of 0 to 1.

The figure shows that most documents that were labelled "A" were grouped together, as indicated by numbers between 0.5 and 1.0 in the upper left block, coloured light grey. The same applies to most texts labelled "B", in the lower right block. Thus the automatic labelling using repeated clustering in two groups agrees to a large extent with the results of the manual determination described above.

However, there are two exceptions. Firstly, document 2—representing folio 48c and supposedly written in hand A—seems to belong to group B, given its high frequency numbers in the dark grey upper right and lower left blocks. Secondly, document 9 (hand B, f. 46c) does not seem to belong to either group A or B, as is demonstrated by



Hand A	Hand B	Contents
	44abcd	
	45abcd	
	46a	
46b		*Aultres balades*
	46cd	
	47abcd	
	48a	
48bcd		
49abcd		*Une complainte amoureuse*
50abcd		*Encore aultres balades*
51a		
	51bcd	*Epistre au dieu damours*
	52abcd	

Figure 5. Distribution of hands in quire 6 of the Queen’s MS. Folios 50bis and 50ter are not mentioned here, because they are blank.

the frequency numbers around 0.5 in both lower blocks. It must therefore be concluded that both documents are different from the other documents in their supposed group only regarding the relation between the direction and the width of the ink trace.

There are many possible explanations of this deviation. Obviously, it is possible that the initial identification of the scribal hands in the documents was not correct. Although the evidence presented by the *Quill* results is strong, a manual codicological examination of both folios in the MS Harley 4431 has yielded no evidence which corroborates this claim. A thorough palaeographical analysis of the aforementioned quire 6—taking into account specimens of every text column—may shed new light on this interesting case. A second explanation could be that the tip of the quill used by the scribe was in a different condition, possibly caused by trimming the pen. A further issue underlying the occurring deviation might be that the scribe returned to his work in the scriptorium after an (extended) break. It is well-known that a scribe’s hand needs to ‘warm up’ before being able to execute its regular, flowing style.

The results generated by *Quill* provide a new, fresh insight into the questions and issues surrounding the scribal hands that are to be distinguished in the *Queen’s Manuscript* in general and in its sixth quire in particular. It is greatly reassuring to observe that *Quill*’s classification of the major part of the documents matches perfectly with those reached through a traditional palaeographical analysis.

6 Conclusion

The palaeographical tests carried out on quire 6 of the Harley MS by means of a computer-enhanced ‘traditional’ method and by using *Quill* not only yield strikingly

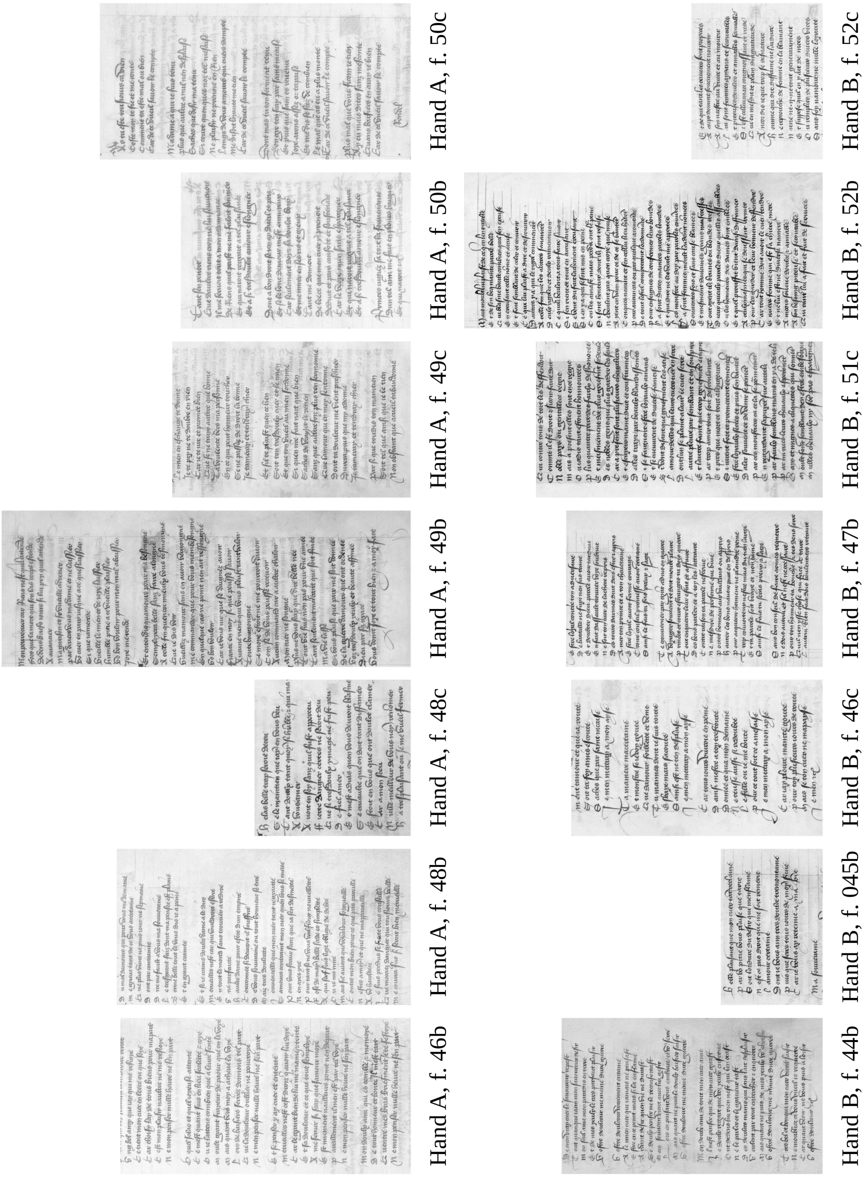
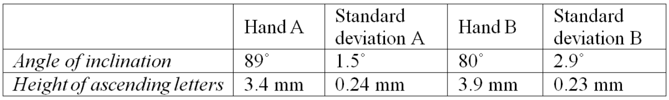
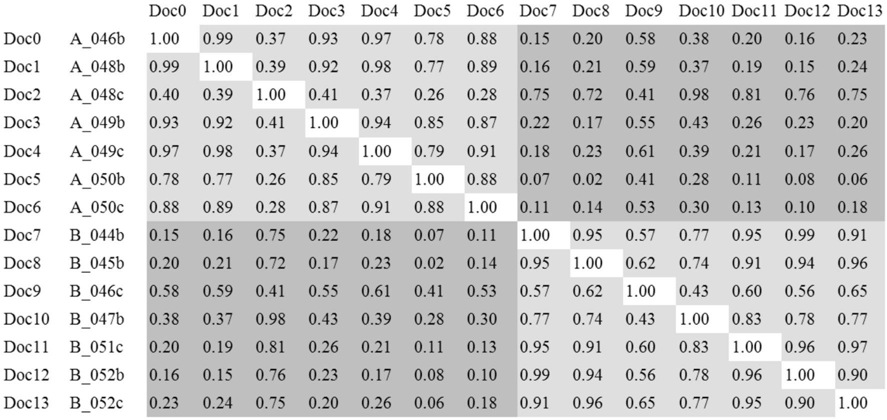

Hand A, f. 46b
Hand A, f. 48b
Hand A, f. 48c
Hand A, f. 49b
Hand A, f. 49c
Hand A, f. 50b
Hand A, f. 50c
Hand B, f. 44b
Hand B, f. 045b
Hand B, f. 46c
Hand B, f. 47b
Hand B, f. 51c
Hand B, f. 52b
Hand B, f. 52c

Figure 6. Overview of the fourteen samples used in the case study. © British Library Board. All Rights Reserved. MS Harley 4431.



	Hand A	Standard deviation A	Hand B	Standard deviation B
Angle of inclination	89°	1.5°	80°	2.9°
Height of ascending letters	3.4 mm	0.24 mm	3.9 mm	0.23 mm

Figure 7. Results of the palaeographical analysis of quire 6 of the Queen's MS, including the corresponding standard deviations.



		Doc0	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10	Doc11	Doc12	Doc13
Doc0	A_046b	1.00	0.99	0.37	0.93	0.97	0.78	0.88	0.15	0.20	0.58	0.38	0.20	0.16	0.23
Doc1	A_048b	0.99	1.00	0.39	0.92	0.98	0.77	0.89	0.16	0.21	0.59	0.37	0.19	0.15	0.24
Doc2	A_048c	0.40	0.39	1.00	0.41	0.37	0.26	0.28	0.75	0.72	0.41	0.98	0.81	0.76	0.75
Doc3	A_049b	0.93	0.92	0.41	1.00	0.94	0.85	0.87	0.22	0.17	0.55	0.43	0.26	0.23	0.20
Doc4	A_049c	0.97	0.98	0.37	0.94	1.00	0.79	0.91	0.18	0.23	0.61	0.39	0.21	0.17	0.26
Doc5	A_050b	0.78	0.77	0.26	0.85	0.79	1.00	0.88	0.07	0.02	0.41	0.28	0.11	0.08	0.06
Doc6	A_050c	0.88	0.89	0.28	0.87	0.91	0.88	1.00	0.11	0.14	0.53	0.30	0.13	0.10	0.18
Doc7	B_044b	0.15	0.16	0.75	0.22	0.18	0.07	0.11	1.00	0.95	0.57	0.77	0.95	0.99	0.91
Doc8	B_045b	0.20	0.21	0.72	0.17	0.23	0.02	0.14	0.95	1.00	0.62	0.74	0.91	0.94	0.96
Doc9	B_046c	0.58	0.59	0.41	0.55	0.61	0.41	0.53	0.57	0.62	1.00	0.43	0.60	0.56	0.65
Doc10	B_047b	0.38	0.37	0.98	0.43	0.39	0.28	0.30	0.77	0.74	0.43	1.00	0.83	0.78	0.77
Doc11	B_051c	0.20	0.19	0.81	0.26	0.21	0.11	0.13	0.95	0.91	0.60	0.83	1.00	0.96	0.97
Doc12	B_052b	0.16	0.15	0.76	0.23	0.17	0.08	0.10	0.99	0.94	0.56	0.78	0.96	1.00	0.90
Doc13	B_052c	0.23	0.24	0.75	0.20	0.26	0.06	0.18	0.91	0.96	0.65	0.77	0.95	0.90	1.00

Figure 8. Frequencies of same-group occurrences after clustering into two groups 10,000 times, based on *Quill.* The numbers show how often each pair of two texts was automatically clustered in the same group, on a scale of 0 to 1. High numbers indicate that the documents were probably written by the same hand.

similar results, but also demonstrate the successful implementation of computer software within the field of palaeography. On the one hand, the case study proves that the traditional palaeographical methods of analysing scribal hands can greatly benefit from the use of computer software which contains built-in measurement tools. On the other hand, we have demonstrated that *Quill*—which is being continuously improved—provides palaeographers with an interesting new approach to distinguishing between scribal hands.

The approach taken by the developers of *Quill* is innovative not only because of its use of the computer as a true aid to the experienced palaeographer, but also because of the way the program analyses medieval handwriting. Of the differentiators mentioned in section 2 of the present article, none corresponds exactly with *Quill*'s approach of using the relation between the width of the ink trace and its direction as a discriminating factor in the search for scribal hands. This new differentiator makes an extremely valuable contribution to palaeography.

Bibliography

Aussems, Johannes F. A. *Christine de Pizan and The Scribal Fingerprint—A Quantitative Approach to Manuscript Studies.* Utrecht: unpublished Research Master's Thesis, 2006. <http://igitur-archive.library.uu.nl/student-theses/2006-0908-200407/UUindex.html>.

Aussems, Mark. "De scribal fingerprint op de tekentafel." *Madoc. Tijdschrift over de Middeleeuwen* 21:4 (2007): 203–13.

Aussems, Mark. "Christine de Pizan et la main X: quelques questions." *Désireuse de plus avant enquerre... Actes du VIe Colloque international sur Christine de Pizan (Paris, 20-24 juillet 2006).* Eds. L. Dulac, A. Paupert, Chr. Reno, and B. Ribémont. Paris: Champion, 2008. 209–20.

Bischoff, Bernhard. *Paläographie des römischen Altertums und des abendländischen Mittelalters.* Berlin: Schmidt, 1979.

Brink, Axel A., Jinna Smit, Marius L. Bulacu, and Lambert R. B. Schomaker. "Quill Dynamics Feature for writer identification in historical documents." [Forthcoming, 2010]

Bulacu, Marius L. and Lambert R. B. Schomaker. "Text-independent Writer Identification and Verification Using Textural and Allographic Features." *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Special Issue - Biometrics: Progress and Directions 29:4* (2007): 701–17.

Bulacu, Marius L. and Lambert R. B. Schomaker. "Automatic handwriting identification on medieval documents." *Proceedings of 14th International Conference on Image Analysis and Processing (ICIAP, Modena, 11–13 September 2007).* Los Alamitos: IEEE Computer Society, 2007. 279–84.

Burgers, Jan W. J. *De paleografie van de documentaire bronnen in Holland en Zeeland in de dertiende eeuw.* Leuven: Peeters, 1995.

Ciula, Arianna. "Digital palaeography: using the digital representation of medieval script to support palaeographic analysis." *Digital Medievalist* 1.1 (2005). <http://www.digitalmedievalist.org/article.cfm?RecID=2>.

Dain, Alphonse. *Les manuscrits.* Paris: Les Belles Lettres, 1949.

Delaissé, Léon M. J. "Towards a history of the medieval book." *Codicologica 1: Théories et principes.* Eds. A. Gruys, and J. P. Gumbert. Leyden: Brill, 1976. 75–83.

Delisle, Léopold V. *Le Cabinet des Manuscrits de la Bibliothèque impériale.* 3 vols. Paris: Imprimerie Impériale, 1868–81.

Delisle, Léopold V. *Mélanges de paléographie et de bibliographie.* Paris: Champion, 1880.

Derolez, Albert. *The palaeography of Gothic manuscript books: from the twelfth to the early sixteenth century.* Cambridge/New York: Cambridge University Press, 2003.

Gilissen, Léon. *L'expertise des écritures médiévales: recherche d'une méthode avec application à un manuscrit du XIème siècle: le lectionnaire de Lobbes (Codex Bruxellensis 18018).* Ghent: Story-Scientia, 1973.

Hardy, Huub J. J. and Wil P. F. Fagel. "Methodological aspects of handwriting identification." *Journal of Forensic Document Examination* 8 (1995): 33–69.

Hindman, Sandra. "The Composition of the Manuscript of Christine de Pizan's Collected Works in the British Library: A Reassessment." *The British Library Journal* 9 (1983): 93–123.

Laidlaw, James C. "Christine de Pizan—An Author's Progress." *Modern Language Review* 78 (1983): 532–550.

Laidlaw, James C. "Christine de Pizan—A Publisher's Progress." *Modern Language Review* 82 (1987): 35–75.

Laidlaw, James C. *The Date of the Queen's MS (London, British Library, Harley 4431).* <http://www.pizan.lib.ed.ac.uk/harley4431date.pdf>.

Mabillon, Jean. *De re diplomatica libri VI in quibus quidquid ad veterum instrumentorum antiquitatem, materiam, scripturam et stilum, quidquid ad sigilla, monogrammata, subscriptiones ac notas chronologicas [...] pertinet, explicatur et illustratur.* Paris: L. Billaine, 1681.

Mallon, Jean. *Paléographie romaine.* Madrid: Instituto Antonio de Nebrija, 1952.

Michel, Lothar. *Gerichtliche Schriftvergleichung: eine Einführung in Grundlagen, Methoden und Praxis.* Berlin/New York: de Gruyter, 1982.

Montfaucon, Bernard de. *Palæographia Græca, sive De ortu et progressu literarum Græcarum, et de variis omnium sæculorum scriptionis Græcæ generibus.* Paris: Guerin, 1708.

Ouy, Gilbert and Christine M. Reno. "Identification des autographes de Christine de Pizan." *Scriptorium* 34 (1980): 221–38.

Schomaker, Lambert R. B., Marius L. Bulacu, and Katrin Franke. "Automatic writer identification using fragmented connected-component contours." *Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition (IWFHR).* Eds. F. Kimura and H. Fujisawa. Los Alamitos: IEEE Computer Society, 2004. 185–90.

Toustain, Charles-François and René-Prosper Tassin. *Nouveau traité de diplomatique, où l'on examine les fondemens de cet art: on établit des règles sur le discernement des titres, et l'on expose historiquement les caractères des bulles pontificales et des diplomes donnés en chaque siècle.* Paris: Desprez, 1750–65.

Traube, Ludwig. *Vorlesungen und Abhandlungen.* München: Beck, 1909–20.

Wattenbach, Wilhelm. *Das Schriftwesen im Mittelalter.* Leipzig: Hirzel, 1871.

Computer-Aided Palaeography, Present and Future*

Peter A. Stokes

Abstract

The field of digital palaeography has received increasing attention in recent years, partly because palaeographers often seem subjective in their views and do not or cannot articulate their reasoning, thereby creating a field of authorities whose opinions are closed to debate. One response to this is to make palaeographical arguments more quantitative, although this approach is by no means accepted by the wider humanities community, with some arguing that handwriting is inherently unquantifiable. This paper therefore asks how palaeographical method might be made more objective and therefore more widely accepted by non-palaeographers while still answering critics within the field. Previous suggestions for objective methods before computing are considered first, and some of their shortcomings are discussed. Similar discussion in forensic document analysis is then introduced and is found relevant to palaeography, though with some reservations. New techniques of "digital" palaeography are then introduced; these have proven successful in forensic analysis and are becoming increasingly accepted there, but they have not yet found acceptance in the humanities communities. The reasons why are discussed, and some suggestions are made for how the software might be designed differently to achieve greater acceptance. Finally, a prototype framework is introduced which is designed to provide a common basis for experiments in "digital" palaeography, ideally enabling scholars to exchange quantitative data about scribal hands, exchange processes for generating this data, articulate both the results themselves and the processes used to produce them, and therefore to ground their arguments more firmly and perhaps find greater acceptance.

Zusammenfassung

Das Forschungsfeld der »Digitalen Paläographie« hat in den letzten Jahren verstärkte Aufmerksamkeit erfahren; zum Teil weil die Paläographen in ihren Urteilen subjektiv zu sein scheinen oder weil sie ihre Argumentation nicht offen legen (können), so dass eine Gruppe von Autoritäten entstanden ist, deren Meinungen außerhalb der Diskussion stehen. Eine Antwort auf diese Situation ist der Versuch, die paläographischen Argumente quantitativer zu machen. Dieser Zugang wird jedoch durch die Mehrheit der

[*] I wish to thank the Leverhulme Trust and the Isaac Newton Trust for their financial support, without which this research would not have been possible.

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 309–338.

Fachgemeinschaft nicht akzeptiert, die unter anderem argumentiert, dass Handschrift per se nicht quantifizierbar sei. Der Beitrag untersucht deshalb, wie man paläographische Methoden objektiver und so somit auch von Nicht-Paläographen leichter akzeptierbar gestalten kann, ohne dabei eine fachliche Diskussion auszuschließen. Zunächst werden ältere Vorschläge für objektive Methoden aus der Zeit vor dem Computer und einige ihrer Defizite diskutiert. Im Anschluss daran wird die forensische Schriftanalyse diskutiert, wobei der Beitrag zu dem Ergebnis kommt, dass ihre Methoden (mit Einschränkungen) auch für paläographische Forschung nutzbar sind. Schließlich werden die neuen Techniken einer »Digitalen Paläographie« vorgestellt. Sie haben sich in der forensischen Schriftanalyse als erfolgreich erwiesen und setzen sich dort immer mehr durch, während sie in den Geisteswissenschaften noch nicht die gleiche Akzeptanz gefunden haben. Der Beitrag diskutiert die Gründe hierfür und macht einige Vorschläge, wie die Software verändert werden müsste, um auch hier größere Akzeptanz zu finden. Abschließend wird ein Prototyp eines Frameworks vorgestellt, der eine gemeinsame Basis für Experimente in »Digitaler Paläographie« bereitstellen soll; idealerweise indem er Forschern die Möglichkeit gibt, quantitative Daten über Schreiberhände auszutauschen, Austauschprozesse für die Erzeugung dieser Daten organisiert, Ergebnisse und die ihnen zu Grunde liegenden Verfahren darstellt und so die paläographischen Urteile besser nachvollziehbar macht und ihnen vielleicht eine größere Akzeptanz verschafft.

1 The Problem: How Many Scribes?

Three questions are commonly asked of palaeographers about medieval handwriting: “when was this written”, “where was this written”, and “were these different things written by the same person.” All three questions are extremely important when working with manuscripts, and all three require slightly different approaches and methods to answer, but the focus of this discussion will be on the third. Many research projects depend more or less explicitly on scribal identification. For example, three major projects are running now on glossed manuscripts from Anglo-Saxon England or early medieval Ireland: the Boethius Commentaries project led by Malcolm Godden, the Irish Glosses project lead by Paul Russell, and the Manchester database of scribes and spellings and its successor, lead by Don Scragg, and all of these depend to some extent on identifying glossing hands (Scragg et al.; Russell et al.; Godden et al.). However, glosses appear in a wide range of scripts and hands, and recognising which glosses were by the same scribe is important but extremely difficult. Even more challenging is identifying where main text and glosses were written by the same scribe or scribes, since glosses so often show different letter-forms, proportions, and aspect because of the different circumstances in which they were written (Stokes 2004; Stokes 2005 Ch. 5). Another example, this one of fundamental importance to late Anglo-Saxon history, is Cotton Tiberius B.iv, the "D" version of the Anglo-Saxon Chronicle. This manuscript contains yearly entries of great historical interest and was written some time between the mid-eleventh century and the early twelfth, probably by a number of scribes. Our understanding of the manuscript and the accuracy of the text depends crucially on whether scribes wrote up the annals year-by-year as they happened or in one block some time after the fact, and to decide this depends utterly on our knowing precisely when the scribes changed. However, at least six different opinions on this subject can be found in print, and there is little if any agreement which is correct (Plummer; Keller; Howorth; Ker; Cubbin; Guimon). These examples are drawn from late Anglo-Saxon England, but the same questions arise again and again in other contexts, and much of medieval studies depends on these questions, whether in the context of history, literature, language, and so on. However, it is difficult at best to answer these questions, and the answer is very often uncertain, something that is not always admitted. A survey of the literature often reveals widely varying opinions even among palaeographers, and that is if the palaeographers even give a firm opinion: in many cases they produce frustratingly vague statements instead, particularly with difficult cases like the D-chronicle mentioned above, or glossed manuscripts, regarding one of which I myself once wrote that "[a]lthough ten or more scribes may have written vernacular glosses in this manuscript, most of the glosses can probably be attributed to just two or three" (Stokes 2005 2:323; compare also Ker). Bernhard Bischoff wrote that "[t]he definite establishment of the identity of medieval scripts in more than one manuscript, or even the establishment of several different hands within a single codex, can [...] be made rather difficult" (1990 44–5). Colleen Sirat took this further when she wrote that "[i]t is obvious that one cannot prove that two texts were penned by the same hand. The only way to persuade other people that this is so is to show them, to give them the feeling that it is the same hand" (2006 493). If Sirat is right then we as medievalists have a problem, since this implies that much of our work depends on nothing more certain than our feelings, and one may well ask how such a discipline can claim to be academic. But it remains to be asked if she really is right, and, if so, what can (or should) be done about it.

2 The Need for Quantitative Palaeography

The issue has been discussed actively since the 70s, if not before, the problem being that palaeographers tend to express qualitative opinions rather than objective arguments and to issue pronouncements that cannot be debated or engaged with meaningfully.[1] To some extent this is necessary, as very little can be argued with certainty in the hu-

[1] At the risk of unfairly highlighting only a few of the very many examples, see the commentaries of the Oxford Palaeographical Handbooks (Bishop 1961; Bishop 1971; Wright) and others such as Bischoff.

manities and particularly in early medieval studies. Furthermore, no argument is ever truly objective insofar as it must depend on assumptions about the data and the interpretation of evidence (Sculley and Pasanek). Nevertheless, palaeography has perhaps been accused of vagueness and subjectivity more often than most other disciplines, and the accusations often seem justified. This is not to doubt the opinions of highly skilled and experienced experts, but rather, as Albert Derolez has noted, that "the method applied hitherto in palaeographical handbooks has produced an authoritarian discipline, the pertinence of which depends on the authority of the author and the faith of the reader" (9). Although in a different context, David Ganz has similarly argued that "the evidence for dating manuscripts must be explained, so that we can learn how a problem may be defined and resolved" (18). However, there remains a methodological problem of how to articulate palaeographical arguments in objective ways; in Derolez's words again, "[h]ow is it possible to proceed in such a way that the description of a specimen of handwriting is as clear and convincing to its reader as it is to its author?" (7) This is not trivial, part of the problem being that subjective impressions are inherently difficult to communicate and cannot be engaged with effectively. But Derolez has proposed a more specific method, namely that "by replacing qualitative data by quantitative ones [...] there is very much to be said in favour of a quantitative approach to a matter so difficult to treat adequately with other techniques" (7–8). Derolez was not the first to suggest this and there had already been significant debate before he wrote, with some scholars protesting that handwriting is inherently fluid and "human" and therefore cannot ever be quantified (Costamagna et al.; Gumbert 1998; Pratesi). Although this possibility must be acknowledged, it has not yet been demonstrated, and indeed those studies which have taken place suggest on the contrary that modern handwriting can indeed be quantified and measured with some significant success (Section 3.3). It therefore seems reasonable to ask whether the same applies to medieval handwriting and, if so, then how digital tools can help.

3 How to Go About It? Theoretical Overview

What is needed, then, is an objective, quantitative method of representing and describing handwriting, of analysing the similarities and differences between scribal hands, and of judging (or at least arguing) whether or not two or more stints of handwriting were written by the same scribe. But is this even possible? As noted in the previous section, scholars have been arguing about this for years, and various methodologies have been proposed. Before turning to the possibilities that computers bring, it is worth first surveying some of the methods that have already been tried, to see how these lessons might be brought into the so-called "digital age."

3.1 Objective Criteria in Palaeography

A useful starting-point for this is a dissertation on the “scribal fingerprint” of Cristine de Pizan which includes detailed discussion of several scholarly criteria for the identification of scribal hands (Aussems). One of the earliest listed there is that of Jean Mallon who proposed seven points for analysis (22–3; Aussems 53):[2]

1. Form, “the morphology of the letters.”
2. Pen angle (*l’angle d’écriture*) “in relation to the base line.”
3. Ductus, “the sequence and direction of a letter’s different traces.”
4. Modulus, the proportions of the letters.
5. Weight, “the difference in thickness between the hair lines and the shadow lines.”
6. Writing support.
7. Internal characteristics, “the nature of the text.”

Other lists since then are similar but tend to specify more criteria and to demand increasing levels of detail. One example is that proposed by Jan Burgers (1:501; Aussems 56–67):

1. Slant.
2. Writing angle.
3. Weight.
4. Modulus.
5. Format.
6. Width of the margins.
7. Ruling and irregularities of the base line.
8. Flourishes and other decoration.
9. “Text structure”, punctuation and use of majuscules and capitals.
10. Abbreviations.
11. Cursiveness between letters.
12. Cursiveness within letters.
13. Characteristic letter forms.

Aussems himself essentially followed this list, although he omitted numbers 5, 6, and 9 as irrelevant for his case, and also number 10 for reasons that will be discussed shortly (70–78). A similar list has been proposed by Michelle Brown in her study of the Book of Cerne (25–26):

1. Aspect, “the overall appearance of the script.”
2. Ductus.

[2] Aussems has also referred to the criteria established by Lothar Michel but these are for modern handwriting and require measurements such as the speed and pressure of the pen which are not applicable to medieval manuscripts. See Aussems 54–6, citing Michel 237–61; see also Aussems and Brink in this volume.

3. Pen angle.
4. Weight.
5. Letter-forms.
6. Mannerisms.
7. Orthography.
8. Abbreviations.
9. Punctuation.
10. Textual Apparatus, namely “devices [...] to assist layout and facilitate reading.”

Alexander Rumble has also printed guidelines for distinguishing scribal hands, “the key to [which ...] is the accurate description of the hands involved” (1994 13). His suggested features for description are (13–15):

1. Treatment of ascenders (including proportions).
2. Treatment of descenders (including proportions).
3. Minims (including weight and shading).
4. Letter-forms.
5. Ligatures.
6. Abbreviations.
7. Punctuation.
8. Orthography.

Finally, a relatively early list was provided by a palaeographer but in the context of detecting modern forgeries, but much of it applies equally well to medieval writing (Brown 1993 259–60):

1. Aspect, including shakiness, layout on the page, beauty, clarity, and tidiness.
2. Spacing.
3. Writing angle.
4. Treatment of loops.
5. Modulus.
6. Punctuation.
7. Formation of common words and syllables.
8. Common groups of letters which may differ according to their position in the word.
9. Ligatures between letters and between words.
10. Dotting of **i** and crossing of **t**.
11. Figures, capital letters, and “odd” letters such as **k**, **g**, **x** and **z**.
12. Pairs of similar letters, such as **n** and **u**.
13. Letters which often have more than one form, such as **d** and **e**.
14. Inconsistencies.

This survey of criteria is far from complete, but already several common elements can be seen in all of them. However, even these apparently scientific criteria do not produce a purely objective analysis, not least because they still include qualitative terms such as "aspect", but also because even the more quantitative criteria are often imprecisely defined and impractical to measure. The criteria involve sitting in libraries and taking very many measurements from potentially hundreds of scribes and thousands of letters; they are therefore impractical, and the best-known study to use such an approach was also methodologically flawed (Gilissen; Derolez 8). But more fundamental problems remain. How does one accurately measure pen-angle, for example, particularly as some scribes deliberately altered the angle of the pen when writing? What does "writing angle" really mean, and which strokes should be measured when determining it, particularly if a scribe wrote a very round or inconsistent hand (see further Maarse)? Some scholars have focused on just a few forms or abbreviations which they considered distinctive, an approach which derives ultimately from that pioneered by Ludwig Traube (1907; Brown 1959 363).[3] However, as Michelle Brown and others have observed, there are difficulties with interpreting even these results. How much does a scribe reproduce the punctuation or abbreviations of his or her exemplar? What about the orthography? Or even the distinctive letter-forms? As Bishop noted in 1961, "the more distinctive [the feature], the more easily imitated" (7–8), and letter-forms, mannerisms and abbreviations are the easiest of the imitable. Scribes certainly imitated script other than their own, such as twelfth or thirteenth-century English scribes consciously imitating (and forging) Anglo-Saxon script, and fifteenth-century scribes in England imitating twelfth-century script (Bishop 1961 7–9; Crick; Parkes 2008 142–4). These imitations are usually obvious, but the adoption of just one or two forms is much harder to detect.

Apart from obvious imitation, this problem of how much scribes were incidentally influenced by their exemplars has been raised many times but answered very rarely. One important response is a series of articles by Angus McIntosh, Michael Benskin and Meg Laing who collectively produced a typology of scribal copying with respect to Middle English dialect. They outlined some ways in which scribal interventions in a text can be identified and suggested ways to undo these interventions and recover the forms in the exemplar (Benskin and Laing; McIntosh et al. 1986 1:12–23; McIntosh 1989a; McIntosh 1989b). They have proposed three categories of scribal copying, "*litteratim*" in which the precise orthography of the exemplar is preserved, "translation" where the orthography is altered to match the scribe's own practice, and a mixture of the two. There are also different sub-categories, such as a scribe who began copying *litteratim* but then lapsed into "translation", or "constrained" scribes who generally follow the exemplar but sometimes give their own orthography instead. Perhaps most

[3] Examples include Gumbert 1976; Muir; Davis 1998; Beneš; and McGillivray; compare also the lists of features presented by Brown 1996 52–60; Brown 1993 259–60; and Rumble 1994 14. A similar philosophy has also been taken by Scragg et al., for discussion of which see especially Rumble 2005 221–5, and Rumble 2006 14–16.

significantly, though, they have found that the vast majority of scribes in the Middle English period fall into the category of “translation”; that is, they tend to adapt the spelling of their exemplars to match their own practices.

An approach like this seems very relevant to palaeographical analysis as well, and some possible approaches have already been suggested, particularly the distinction between a scribe’s “graphic” and “linguistic” profile, namely, the handwriting on one hand and the orthography and punctuation on the other (Parkes 1994, esp. p. 24, citing McIntosh 1974 and McIntosh 1975). Specifically, we may ask whether scribal practices in script, punctuation and capitalisation follow the same patterns as in orthography and dialect, and, perhaps more interestingly, whether the methods proposed to recover the dialect-forms of an exemplar can be used to recover the letter-forms as well. Certainly we can find examples of both “*litteratim*” and “translation” in letter-forms. The latter is the norm, insofar as most scribes copied with their own natural script, but we have already seen examples where this was not so. There are also examples of scribes starting with one script and lapsing into another, such as Cambridge, University Library MS Ff.1.23, the so-called “Winchcombe” or “Canterbury” Psalter which was written in the first quarter of the eleventh century, probably at Canterbury (Dumville 1991 40–41; Stokes 2005 1:41–42). This scribe began writing the Latin text of the psalms in a careful Anglo-Caroline script but introduced more and more vernacular letter-forms, apparently by mistake, and after about a page or so gave up entirely and wrote the Latin and Old English in the same English Vernacular minuscule (Stokes 2005 1:68, reproduced by Robinson 2:pl. 17). This example probably says less about the exemplar and more instead about the new requirement to differentiate between languages by script (for which see Bishop 1971, Dumville 1993, and Stokes 2005), and this is one point where orthography and script diverge.[4] Nevertheless script was certainly taught, and there seems often to have been a strong sense that certain manuscripts or even texts should be written in certain scripts (Brown 1993 201–2; Lieftinck 1964 1: xiii–xvii). Furthermore, even vernacular orthography may have been standardised in some places even in the early Middle Ages, such as in England from the late-tenth through to the early-twelfth centuries where it may have been more consistent than the vernacular script was (Gneuss esp. 70; Gretsch 69–83). Similarly, the apparent readiness of scribes to write very different scripts alongside each other in the early eleventh century in England, as demonstrated by many vernacular manuscripts such as the so-called “Beowulf manuscript”, also suggest that the pressure to write a particular style of script may not have been as strong as we might like to think (Stokes 2005; reproduced by Zupitza and Kiernan). Of course to address this question properly requires very much more research, and this is precisely something that databases of scripts and spellings and its

[4] Compare Bishop: “The difference in aspect between the Latin and the vernacular script need hardly be considered [and was at] no time so marked as to disguise a scribe’s identity” (1961 4).

successors should help us answer (Scragg et al.), but the point remains that even the apparently objective criteria given above still require a good deal of interpretation.

There is a further difficulty with the different approaches listed here: although similar, there are important differences between them, and yet there is no clear way of testing them or deciding which should be used, or indeed having any meaningful way of knowing which, if any, lead to valid results. However much one might argue that "all knowledge is situated and contingent" (refuted by Drout §§10–12 and Shippey §26) the fact remains that a surviving manuscript was once written by one or more people at one or more place(s) and time(s) in history, and in this sense handwriting identification has a "correct" answer that palaeographers seek, whether or not it can be attained in practice. However much one might discuss individual strengths and weaknesses in each of the methods discussed above, it is ultimately difficult or impossible to know which is most "correct". Furthermore, the very question of scribal identity depends in turn on the assumption that the handwriting of no two people is the same, and yet this assumption is not normally questioned by palaeographers. To some extent this uncertainty is inevitable, and we should neither demand nor expect that palaeographical results will always be certain. Nevertheless, it is not ideal to have an entire discipline the validity of which has been assumed but not firmly demonstrated.

3.2 Objective Methods in Forensic Document Analysis

Fortunately most of these difficulties have already been raised in forensic document analysis.[5] Forensic document analysts have been critisized for some of the same shortcomings as palaeographers, including the inability to verbalise their methods and the variation in their results, and these uncertainties have even reached national headlines in the United States (Kam Wetstein and Conn 6–7 and 12; Liptak). When forensic document analysts do articulate their methods, furthermore, they seem to follow principles much like those of palaeographers, referring to features such as aspect, slant, writing-angle, shading, cursiveness between and within letters, characteristic letter-forms, and particular idiosyncracies (Kam Wetstein and Conn 12). However, document analysts must withstand cross-examination in court, and so they are forced to provide clear objective arguments. Furthermore, the United States Supreme Court has a recent guiding principle that judges must evaluate expert testimony for factors such as whether or not the method has been tested, what the potential rate of error is, and how reliable (and therefore reproducible) the results are (Srihari et al. 2002 856–7). This guiding principle has twice brought to bear on recent cases involving document analysis, as a result of which the Supreme Court was asked to rule on the objectivity and validity of handwriting identification (US v. Prime; US v. Thornton). New York state justices

[5] Forensic document analysts (FDAs) are also known as questioned document analysts (QDAs), or forensic or questioned document examiners (FDEs or QDEs).

also commissioned a study to determine if the identification of handwriting is objective and if more objective methods are possible, and it is significant for our discussion that "objective" was understood to mean "automatic" and determined entirely by computer (Srihari 2001). As a result, the State University of New York has a very active centre for the study of individuality in modern handwriting, and they have been studying questions of importance to forensic document analysts and palaeographers alike, questions such as whether handwriting is indeed individual (Srihari et al. 2002), how accurately trained and untrained people can identify samples of handwriting written by the same person (US v. Prime 11–12), whether computers can identify writers automatically (Srihari 2001), and even whether and to what degree handwriting varies between twins who have the same education and (presumably) much the same biological mechanisms (Srihari Huang and Srinivasan). Fortunately they found that handwriting is indeed discriminable and that trained experts can correctly identify passages written by the same person with a fairly consistent level of accuracy (US v. Prime 11–12; Kam Wetstein and Conn; Kam Fielding and Conn). They also found that experts are significantly better than untrained people at identifying which samples were written by the same people even without the benefit of any laboratory equipment; more interestingly, they found that "nonprofessionals" tend to produce many more "false positives", that is, untrained people were found generally to underestimate the degree of similarity in different people's handwriting (US v. Prime 11–12; Kam Wetstein and Conn; Kam Fielding and Conn).[6] They also found, again unsurprisingly, that the "handwriting of twins is less discriminable than that of non-twins", and that "error rates with identical twins were higher than with fraternal twins", but that the handwriting of twins can still be identified with an error-rate of about 13%. Perhaps more importantly for our purposes, they developed fully automatic systems which could correctly identify the writer of a given sample 95% of the time or more, and that the success-rate in almost all of these tests was about the same for human specialists and purely automatic systems (Srihari et al. 2002 871; Srihari Huang and Srinivasan 2008).

These are all precisely the questions that have been raised about palaeography, and indeed the relationship between the two fields is being noticed more and more recently (Davis 2007; Stokes 2007/8). However, although modern forensic processes have a lot to teach students of medieval handwriting, there are also important differences between the two. One is that forensic document analysts often can (and ideally must) obtain large samples of the suspect's handwriting, preferably written at different times and

[6] The study in question found that untrained people incorrectly attributed two similar samples to the same person 38% of the time compared to 6.5% of the time for experts, and that they correctly matched documents written by the same person with about the same average accuracy as experts (Kam Fielding and Conn). It is an interesting question how expert palaeographers would fare in controlled tests of (modern) handwriting identification as described by Kam Fielding and Conn and US v. Prime 10, or those provided by CEDAR.

in different circumstances, to build up a full picture of the individual and his or her variation (Davis 2007 255; US v. Prime 11–12). However, the palaeographer rarely has this luxury, and even if a substantial corpus has been attributed to a scribe, it is rare (at least for the early medieval period) for those attributions to be certain. Forensic analysts also rely on features that are not generally available to the palaeographer. For example, the pressure exerted on the pen has been cited as important for forensic analysis, particularly for detecting forgeries, and this can be measured with an electrostatic detection apparatus which detects indentations in paper (Kam Wetstein and Conn 12). However, medieval quills do not require pressure to write in the way that modern pens and pencils do, and parchment does not hold indentations, particularly not for centuries, so electrostatic devices would not yield any information.[7] One also suspects that modern writers are much less practiced than medieval scribes, and therefore that modern handwriting varies much more than medieval does, at least for samples from the same region and period.[8] Indeed, the methods employed by forensic analysts, and particularly the automatic handwriting-identification systems they use, often explicitly exclude skilled forgeries or even imitations, suggesting that they may be inappropriate for medieval script (Srihari et al. 2002 857; Kam Wetstein and Conn 7). All these issues suggest that forensic techniques cannot be applied uncritically to medieval script, and although some techniques of forensic analysis have been successfully trialled on medieval documents, the results require further improvement (Davis 2007; Bulacu and Schomaker 2007; Stokes 2007/8).

3.3 "Digital" Palaeography

Returning to the issue of objective methods in palaeography, the question remains what methods can and should be used, and what we can learn from other related fields such as forensic document analysis. One striking aspect of recent research in forensics is the use of computing in the attempts to quantify the field and particularly to develop objective methods. Such an approach has also been hinted at, although not stated explicitly, by Derolez when he suggested replacing qualitative measurements with quantitative ones (7–8), since quantitative methods now normally imply digital ones. Indeed, most of the approaches discussed in Section 3.1 benefit from the use of computers. Features such as the angle and width of strokes can be measured much more easily with high-quality images than they can from manuscripts; images can easily be magnified if the resolution

[7] One calligrapher I have spoken to, Michael Gullick, has described writing with a pen as pushing ink across the parchment, a process which exerts almost no pressure on the page. Pressure on the quill can sometimes be detected by the strokes that remain, but this is much more difficult and less objective than electrostatic devices.

[8] For some examples of variation see the samples provided by Srihari et al. 2002 857 or CEDAR, and contrast those with plates of medieval script such as those by Watson, Robinson or Lieftinck and Gumbert.

is good enough; examples of letter-forms can be cut out and stored in databases for comparison, and so on (Stokes, Palaeography and the 'Virtual Library'). However, these methods are not new and do not depend on computing, they have just become easier and therefore widespread with the advent of the so-called "digital age". In contrast to this are some entirely new approaches which have emerged just in the last five years or so and which have only become possible with the combination of powerful computers and high-quality digital images in what is starting to be called "digital palaeography" (Ciula; but for a very different use of the term see Hirtle). In essence this is a logical extension of the older methods: it is again taking statistical measures of handwriting and then using these measures to make inferences and quantify similarities and differences between hands. The crucial point is that the earlier methods use statistics which were developed by a person sitting down and doing all the counting and measuring. In contrast, this new approach is fundamentally different: we now take images of handwriting and feed these into a computer, then ask the computer to make comparisons and find out which hands are closest to others (Bulacu and Schomaker 2007; Ciula; Stokes 2007/8; see also Hofmeister et al. in this volume).

The underlying principle is to use techniques in computer science, especially in image-processing and data-mining, to generate statistical measures of handwriting and to use these to compare the handwriting in ways that could never have been done previously. These approaches therefore constitute two stages, the first is sometimes known as "feature extraction" and involves generating the numerical measurements, and the second, "data mining", constitutes finding similarities and classifying handwriting based on these measurements (Stokes 2007/8). One example of feature extraction is to take many examples of a given letter (or ligature) written by a single scribe and generate a composite "average" letter from all of them (Ciula). Alternatively, one might break down every stroke in a sample into thousands or millions of tiny line segments and measure the angle of every such segment, or indeed the angle between adjacent sections. We can also do less obvious things like overlay a sample of writing on top of itself and then slide the top sample across and see how much the writing overlaps; the more regular the hand, the greater the overlap (Bulacu and Schomaker 2007; Stokes 2007/8). One or more of these sets of data, these quantitative measure of features, can then be used to generate a statistical profile of each sample of handwriting, and these profiles are then used to compare different samples and to measure the mathematical distance between them. Neither feature-extraction nor measuring distances needs to be especially complex: software can be written quite easily which does this at a basic level, and forensic document analysts have already tested these methods and have developed systems which can correctly identify the writer in 90–95% of cases or more (Srihari et al. 2002 871; Bulacu and Schomaker 2006 285). Other possible methods are much more complex, however, often rely on postgraduate-level mathematics, and their potential is far from fully exploited (for one possibility see Kingsbury).

However promising this may seem, "digital palaeography" seems to have received almost no acceptance and very little interest from so-called "traditional" palaeographers. This is partly because the technology is not yet mature, but this is not a complete explanation. It may also be because most work in this field to date has involved small groups working for relatively short periods, rather than the large, interdisciplinary groups with extended funding that digital humanities often requires. The System for Palaeographic Inspections, for example, was developed by postgraduate students in computer science and one doctoral student in digital humanities, and was never completed (Ciula §§13–14). Software developed for modern forensic analysts has been applied to medieval writing but apparently without the directly involvement of scholars in the humanities (Bulacu and Schomaker 2007). A project to identify medieval scribal handwriting led by a computer scientist and a palaeographer was announced in 2004, and the UK Arts and Humanities Research Council awarded funding to the principle investigator in 2006 but the results have not yet appeared publicly to my knowledge (Intute; AHDS). Finally, the software to quantify differences in scribal hands described in Section 4 has been developed by one person, the author of this paper, working as both computer scientist and palaeographer, as part of a two-year project funded by the Leverhulme Trust. Although none of these is trivial, and although other projects are now emerging, including some described elsewhere in this volume (see contributions by Hofmeister *et al.*; Aussems and Brink; and Ciula), we have not yet had the large groups with experts in a range of fields, computing and humanities, palaeography, image-processing, data mining, but also interface design, database design, developing XML schemas, and so on, and these interdisciplinary groups are now normally required for work in the digital humanities (Pierazzo).

These difficulties, the relative immaturity and the lack of sustained interdisciplinary research, can both be resolved relatively easily given time and resources. However, another problem is perhaps less obvious but still significant, namely that of understanding and engagement. In some cases of software designed for palaeographical analysis, as indeed for other applications of digital humanities, it is not clear exactly what the computer is doing, either because the particular technique requires a lot of *ad hoc* human intervention which is not properly documented, or because the software is proprietary rather than open-source. One example of this is the image enhancement performed by Fotoscientifica, a company which recovers text from damaged manuscripts using multispectral photography and image enhancement (Fotoscientifica). Their results are spectacularly successful and yet their services are sometimes not used because of concern about the degree to which they enhance and the lack of openness about what they have enhanced and how they have done it (Craig-McFeely 2007/8 §§62–3).[9] Even if the

[9] I myself encountered such concerns among colleagues when working for the British Library on the *Rinascimento Virtuale* project to recover Greek palimpsests in 2003–4.

methods are fully documented and reproducible, however, and even if they are communicated clearly using recognised standards and terminology, scholars still require a good understanding of many complex fields to fully appreciate and engage with the results. Indeed, this concern has already been raised explicitly by Tom Davis in a footnote on computing in palaeography. He noted that "these methods are unlikely to replace, though they may supplement, the work of the document analyst, because, however powerful computers will (surely) become, it will probably not be possible to cross-examine them" (266 n. 27). Similarly, researchers in forensic document analysis have argued that "black box" answers rather than verbal reports have contributed to juries tending not to accept automatic methods (Schomaker 2007 §6), and much the same has recently been said for data-mining in literary criticism (Sculley and Pasanek 421). This is by no means to underestimate the ability of medievalists to move between disciplines and to grasp very complex concepts outside their main field. But it does seem fair that we as medievalists in general and palaeographers in particular cannot be expected to understand the intricacies of postgraduate-level mathematics and computer science, and if we cannot understand them then we cannot evaluate them properly or debate their results. We therefore have the same authoritarian discipline as before, with final pronouncements that must be either accepted or rejected wholesale. The difference is that the authority is now a machine.

On the other hand, the results of "digital" palaeography look very promising and should not be discarded lightly; as noted in Section 3.2, experiments with modern handwriting have given successful identifications in 95% or even 98% of the time. Indeed, as discussed above, projects today in digital humanities routinely involve large groups of experts who cannot fully understand each other's fields and who have to trust their validity to some extent. One might even cite the precedent of digital methods such as genetic algorithms which have been used to design new and very effective electronic circuits even though the engineers are sometimes surprised by the results and cannot always explain how they work (Rahmat-Samii and Michielssen 1999, especially 245–6 and 272–7). Nevertheless the computer is a tool to aid us, and like any tool it must be understood before it can be used correctly. In this respect palaeography is not like electronic engineering, or indeed like some branches of digital humanities, insofar as the engineers (and some digital humanists) can test the results of their algorithms, and as long as the results are valid then the details of how they were obtained are not important. However, as has been stated several times already, palaeographers cannot easily test their results, computer-generated or otherwise, and even if they can then those results do not necessarily hold when applied to different scripts or different types of manuscripts, since the methods depend on assumptions that may or may not still be valid. If we do not understand the algorithms, though, then we cannot know on what assumptions our algorithms depend, and therefore we cannot know if they still hold in the new situation. This uncertainty means that our tests are no longer useful.

With this in mind it follows that we *must* be able to "cross-examine" the computer, to use Davis' phrase; that is, we must program the computer in such a way that it is cross-examinable. In other words, even if we as palaeographers or medievalists cannot readily understand the method we should still be able to interpret the results. Rather than having a computer announce that Hand A and Hand B are by the same scribe, it seems much more useful for it to state that Hand A and Hand B both have an average inclination of X°, and an average proportion of width to height of Y, and ascenders of relative length Z, and so on. This sort of meaningful information is perhaps more likely to be trusted than vast quantities of meaningless data or electronic pronouncements of scribal identity; as noted above, this point has also been suggested for forensic document analysts and data-mining in literary criticism. It has recently been argued that systems for forensic document analysis should present information in verbal reports, including margins of error in their results, if their results are to be accepted by juries (Schomaker 2007 §6), and we may reasonably argue the same for medieval handwriting and palaeographers as well. One may well argue further that the computers should not even try to judge scribal identity, but instead that they should present data for experts to interpret; either way, though, it seems useful for that data to be intelligible and ideally to give new insights into ways of seeing and comparing handwriting. This is perhaps a more beneficial way for computers to be used in palaeography. This also suggests renewed scope for the old style of quantitative methods now that we can use computers to organise and process our data. The difference now is that we can handle a much large volume of data than before by using databases and spreadsheets in what could be called "statistical" or perhaps "computer-aided" rather than purely "digital" palaeography.

3.4 Suggestions for a Successful System

Now that the background and previous attempts have been considered, it is worth asking what criteria are necessary for a computer-based system of handwriting identification to be as successful as possible while still being acceptable to palaeographers and medievalists. The first criterion to emerge from this discussion is that whatever is done should be reproducible; this is a basic criterion for acceptability in the scientific world, and reproducing the results of lab experiments in the sciences is considered valid (and necessary) research in its own right. Such reproduction is not considered valid research in the humanities, but it has been recommended for data-mining in literary criticism and for digital humanities more generally (Sculley and Pasanek 423). Similarly, a judge in the US Supreme Court has criticised one set of studies on handwriting identification precisely because the data was not released and so the work could not be verified (US v. Prime 12). Even though it is unlikely that anyone will reproduce a long and detailed study of handwriting *in toto*, it still should be possible in principle to reproduce the experiment and verify its accuracy, otherwise the authority of the study will again be dependent entirely on the person who produced it.

As well as being reproducible, the process should also be debatable; that is, it should be possible to understand and evaluate the assumptions which underlie the analysis and each of the stages used in getting to a result. This also suggests that the results should not be final but should themselves allow for (human) interpretation, or at least understanding, and should also indicate the mathematical level of confidence in the result. Sculley and Pasanek have recently demonstrated that interpretability is not a sufficient criterion for evaluating results in data mining (421), but I would suggest that it is a necessary one. This is another aspect of "communicability" but applying to the outcome rather than the process. The computer should generate evidence which can be "cross-examined" and interpreted by us, the scholars, rather than producing the impenetrable answer of a final authority.

In practice, these criteria imply that the process should be documented and made open in a way that can be communicated effectively and understood by those in the field (compare Pitti 482). This is rarely the case in studies to date, and is particularly rare in computer-based work in manuscripts. To take a slightly different example, some scholars (including myself) use Adobe Photoshop or the GIMP to enhance images of damaged manuscripts and thereby recover lost readings (Craig-McFeely and Lock; Stokes, Recovering Anglo-Saxon Erasures). However, it is very difficult indeed to record precisely what enhancements are done on a particular image, and this in turn makes it difficult for anyone else to verify the results. Furthermore, proprietary software is by definition opaque, and any documentation is usually platform-dependent and unsustainable in the longer term. Thus, to use the example of Photoshop once again, even if one carefully notes every minute enhancement that one makes, this information is still only useful if someone else has exactly the same version of Photoshop and will almost certainly be useless in a few years time when the software has changed and the old version is no longer available. Furthermore, it is not at all clear exactly what software like Photoshop does in particular cases, and so it is difficult or impossible for anyone else to evaluate. On the other hand, if one uses an open standard such as METS to record the precise details of all the algorithms employed then this can be interpreted by anyone else with the required skills, it can be repeated in future, potentially with other software, and it is not tied to a single version of a single application. In this way the criteria to be reproducible and debatable imply a further one, namely to be communicable: a system should allow the entire process to be documented, preferably automatically and without the user having to intervene, and employing open standards for information interchange (Stokes, Recovering Anglo-Saxon Erasures).

In addition to these so-called "digital" aspects, there are also some "humanities" requirements. Certainly any system must allow for a lot of scribal variation. Manuscripts are written by people, not machines, and people change according to many different factors, so this must be taken into account by any palaeographical method, digital or otherwise (Bishop 1961 4–9; Costamagna et al.; Gullick 23). Nevertheless, the entire field of palaeography (and forensic document analysis) is based on the assumption that everyone's handwriting has some deep, innate and inherently individual characteristics which do not vary, or at least vary slowly and can be documented.

So our hypothetical method can (and surely must) assume some underlying commonality in handwriting, although it is by no means clear what kinds and degree of variation this should entail. A successful method must therefore not be too rigid and must accommodate this flexibility, ideally allowing the user to program how much and what sorts of variation she or he has in mind. Indeed flexibility seems to be the key, not least at this relatively early stage when there is still so much uncertainty about what methods will work best and how the software should be developed.

As well as flexibility in allowing variation of handwriting, a system such as this should also allow for flexibility in the methods used for generating the measurements. As discussed in Section 3.3, there are quite a number of competing algorithms which are already being used by forensic document analysts, but it is by no means clear which one, or which combination, will be most successful for medieval documents. Indeed, it seems entirely possible that different types of document will respond better to different combinations of algorithms, and Sculley and Pasanek have advocated always using different methods for data-mining in literary studies (423). Furthermore, many methods have been developed already which have not yet been applied to medieval handwriting but which will probably be useful in this context, and new methods will continue to develop (for one promising example see Kingsbury). Therefore, any software designed to analyse handwriting must allow users to easily add new functionality, otherwise it will quickly become obsolete. Rather than providing a fixed process for analysing handwriting, it seems much more useful at this early stage to provide a common framework which allows researchers to test different methods in a consistent way, allowing one to compare the different results. Just as it is difficult to assess the efficacy of methods in "traditional" palaeography, so also is it hard to comparing digital methods. For two different algorithms to be usefully compared, they must be run in the same circumstances with exactly the same images, the same "correct" classification, and so on, and this assumes that the required outcome is already known. These conditions can potentially be achieved but only if all researchers release the data-sets that they used to test their systems, and the more that this requirement is built into the system the better.

4 A Practical Suggestion: The Hand Analyser

Now that these criteria have been presented, it remains to ask how they might be put into practice. To this end software has been developed to implement a framework for the analysis of scribal hands. As discussed in Section 3.3, this has been developed entirely by the author of this paper as part of a two-year project to investigate objective methods in palaeography. Being a "lone scholar" working on all aspects of the topic, theoretical and practical, "digital" and "humanities", has necessarily limited the amount that can reasonably be achieved, and the framework is certainly not considered to be final or even necessarily usable by palaeographers in its present state. Instead, it is designed to be a platform for testing the methods and principles that would ultimately form part of such a tool for palaeographers and scholars in the humanities more generally.

4.1 Design Principles

Several design decisions follow from the principles outlined in Section 3.4. The requirement to provide a common framework for disparate scholars to share information and test each other's results, along with the basic requirements for sustainability, make Java the obvious choice of programming language. Software written in Java is multi-platform by nature and (in principle, at least) should work on future computers without needing to be rewritten or recompiled. Furthermore, Java now has a very large number of standard libraries of pre-existing code which can be incorporated into any new software, and it has such a wide user-base that these libraries are unlikely to be discontinued for some time, if at all. In particular, the Java Advanced Imaging library (JAI) provides a lot of useful functionality for image-processing. This is a standard library which can be distributed freely and which ships with most installations of Java; it is therefore already installed on most computers and can be freely downloaded from the Java website if not. The source code is under licence to Sun Microsystems but is "open" and may be modified for research use.[10]

Perhaps the most fundamental is the requirement for extensibility and the easy inclusion of different modules. To this end, the system has been designed as a fully modular framework in which the processing is done almost entirely by plugins, where each plugin runs a process to generate a single set of measurements from a single set of features, probably by implementing one or more algorithms in image-processing.[11] For

[10] The source code is available under the terms of the Java Research Licence for non-commercial use and the Java Distribution Licence for commercial use (JAI-Core). Note that this use of Java is a change from previous work discussed by Stokes 2007/8, for which C++ was used and an imaging library from Delft University. Not only was this older software platform-dependent but the imaging library was discontinued between the first development of the software in 2004 and the beginning of the Leverhulme fellowship in 2007. A "second generation" has now continued support and released a version of the library for MacOS X but this still cannot be distributed freely and has a much smaller user-base than the JAI and so is in much greater danger of being discontinued again.

[11] This structure was inspired by John Bradley's discussions of Pliny which in turn draws on that of the Eclipse workbench for software development (Bradley; Birsan).

example, the directions of the edges of strokes has been used as one way of measuring handwriting (Bulacu et al.), and so a plugin can then be written to implement this process. Other plugins can implement different processes, and the users can hence choose which processes they want to test on their particular samples and thereby determine the optimum combination for their particular cases, as well as writing and testing their own new processes.

To allow the exchange of information, as well as accommodating practical issues such as the long time that is often necessary to process high-quality images, the system treats "hands" as distinct objects, where each "hand" contains a complete set of information about a given scribal hand. It therefore includes the URL and other relevant metadata of the image, a full record of the processes that have been carried out on that image, and the full set of measurements generated by each process. The system therefore comprises one or more processes that are run in turn on one or more hands, the results of which are then stored by a "hand" object and can be used to measure the statistical distances between the various samples. A "hand" file combined with the necessary plugins and image file is therefore sufficient to reproduce the process of analysis, but the "hand" file alone contains sufficient information to compare it with other scribal hands. This allows the exchange of information and also means that users need not rerun all the plugins on each hand every time they wish to access the data, an important benefit for such an intensive process.[12] Since each "hand" includes a record of which process has been run on it, and the data generated by that process, it follows that any two hands can check which processes they have in common and use the data generated by the common processes to measure the distance between them. For example, Hand A might have had three processes run on it, say Horizontal Runs, Vertical Runs, and Autocorrelation, in which case it will include three sets of measurements, one for each process.[13] Hand B might also have had the Horizontal Runs and Vertical Runs, but then had Edge Directions and Hinge Directions, and therefore contains four sets of data. However, each hand "knows" what has been done to it and "knows" that it is not meaningful to compare data generated by different plugins. A comparison of Hands A and B will therefore use the results of the Horizontal Runs and Vertical Runs and ignore the other sets of data. This then allows different scholars to run processes on different sample images and pool the resulting "hands" and in this way a very large database of scribal hands could be built up by many different scholars contributing their data from around the world.

The framework itself is further divided into two packages, one containing the core modules which drive the system, store the data, and coordinate the plugins, and the

[12] The five processes described by Stokes 2007/8 on a single image of 1370×490 pixels can take approximately 90 seconds on average when running on a MacBook Pro (2.4 GHz Intel Core 2 Duo processor with 2 GB RAM running the Java 2 VRM version 1.5 under MacOS 10.5).

[13] For these terms and those that follow see Stokes 2007/8.

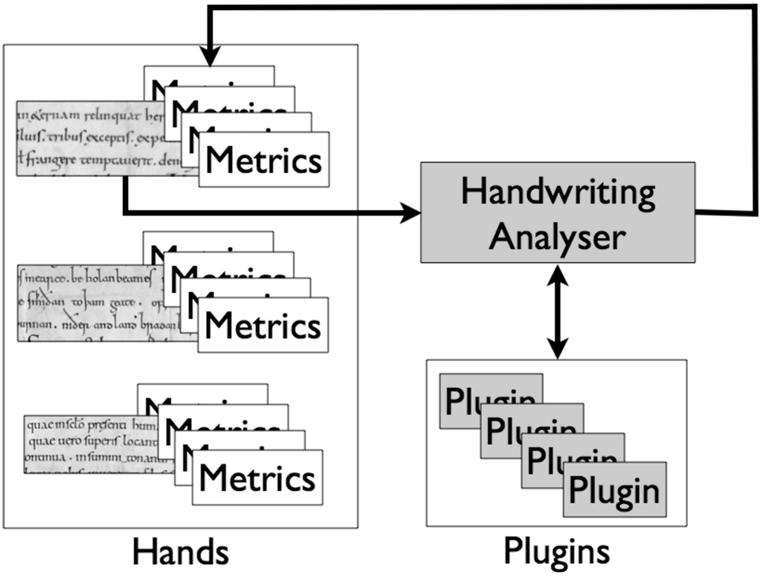


Figure 1. Structure of the Image Analyser.

second providing a graphical user interface (GUI). This allows the GUI to be adapted and expanded independently of the processing and, if desired, to be entirely rewritten, thereby allowing different users to have different interfaces depending on their needs such as, for example, having a desktop application and a web-based service.[14]

At a more basic level, each plugin must automatically log every step taken by the user and allow this log to be exported in MIX/METS format.[15] This is relatively easy to achieve, except of course that one has no way of knowing that all future plugins will conform to this requirement. Instead, everything possible should be done to encourage those writing plugins to conform to this requirement, and specifically the capability for

[14] This is similar to, but different from, the principle that content and presentation should be separated to allow long-term survival of digital resources (O'Donnell 2004 §2; TEI v1). Here, the data can indeed be exported (as is discussed shortly), and the use of Java also makes it more sustainable, as Java is itself separated into different layers and so is inherently platform-independent, and this combined with its very widespread use makes it comparatively sustainable (Gosling and McGilton 1.2.3; compare also O'Donnell 2007 65–6 on the relative longevity of the Java produced by Kiernan).

[15] MIX is an XML implementation of ANSI/NISO Z39.87–2006, §10 of which applies to image processing, and MIX itself is often used to extend the METS standard for technical and administrative metadata (NISO; MIX; METS).

generating such logs is an inherent part of the system's design and should be as easy as possible for authors of plugins to use.

The requirements of reproducibility and sharing of information create a practical difficulty. On the one hand, this requirement implies that images should be stored in the system and circulated as a crucial part of the data; after all, no process can possibly be reproduced if the original images are not available. However, circulating images in any way is not "private use" and therefore violates the terms of use for many digital repositories and libraries (Padfield 165; Case and Green; Stokes, Palaeography and the 'Virtual Library'). For this reason, the system should ideally store URLs to images which are freely available instead of storing the images themselves. As long as the entire process is recorded in suitable detail then users should still be able to reproduce it by using the online images. Unfortunately this has several drawbacks (Stokes, Palaeography and the 'Virtual Library'), not least that the images which are made freely available are often not of sufficiently high quality and are almost always in JPEG format, and this lossy format is not suitable to very demanding applications (Craig-McFeely and Lock; Craig-McFeely).

This requirement of working from publicly available images also has implications for the way in which plugins should be designed. Specifically, they should not assume any preprocessing but should instead begin with the raw image as found at the relevant URL and should log every single step required to produce the output. Indeed the plugins that have been developed to date are all designed to operate on the image as a whole with minimal human intervention. This is in contrast to some other systems in which the user must select portions of the image and often classify those portions, such as selecting all examples of a particular letter, for example. The system being discussed here certainly allows this alternative approach, the slight difficulty being that the sections of images being processed must be carefully defined and recorded in order to comply with repeatability requirement. However, there are now standards for recording this (TEI §11.1) and indeed tools for selecting portions of images and even developing databases of letterforms from images (Holmes; IDP), and in principle these could be turned into plugins for this system.

Another concern relates to the information generated by each plugin and stored in each hand. As discussed in Section 3.4, many of the algorithms used for handwriting identification are different from the principles used by human palaeographers or forensic document analysts, and the algorithms can generate hundreds or even thousands of numbers which are not readily intelligible to human users. It is therefore desirable that plugins present data which can be understood by human users. This need not preclude very large data-sets, as one possibility would be to allow access to the raw data but also provide the mechanism for displaying that data in a human-readable way such as graphically or even interactively.[16] This cannot be enforced, but the structure of the plugins allows the possibility, and indeed the use of Java means that even interactive interfaces can be designed relatively easily and in a way that is platform independent and relatively sustainable (see p. 326).

One example of such an interactive plugin is the Image Enhancement module. This has been designed to address the need to "clean up" images by removing as much of the background as possible and presenting only writing to the computer for analysis, since otherwise the computer can be mislead into interpreting the parchment, folds, stains and so on as ink. Although this can be done easily enough with software such as Adobe Photoshop or the GNU Image Manipulation Program (GIMP), this software has limited value in a scholarly context (Stokes, Recovering Anglo-Saxon Erasures), not least because it does not allow users to easily record the steps taken in the way discussed above, and even if it did this information cannot be readily incorporated into the system for handwriting analysis. For this reason a module was designed which lets the user process the image in a controlled way to remove as much background noise as possible, the objective being to have simply black ink on a white background, and the result of this is documented as part of the overall process of analysis and can be exported, reproduced, and so on. Indeed, the need for a controlled system for image enhancement with automatic documentation is not limited only to handwriting identification but is a desideratum for manuscript studies more generally (Stokes, Recovering Anglo-Saxon Erasures; Craig-McFeely and Lock; Craig-McFeely), and so the module has been designed in such a way that it can also be used as a standalone application.

5 Conclusions

Most of the principles discussed above have been implemented in the prototype at the time of writing, and they are successful insofar as they provide a framework for experimentation, although the framework is nowhere near as sophisticated as that described by Birsan. However, I make no claim that the software is ready to be used by palaeographers or anyone else to establish scribal identity, and it is partly for this reason that no results are presented here (but see Stokes 2007/8). Indeed, as discussed throughout this paper, this system (and probably every other) should not be taken as a "black box" and used uncritically, and its success can only be judged when it is made widely available and tested by the community at large.

On another methodological note, the processes tested in the framework so far all rely on analysing the image as a whole, without the computer having any knowledge of letter-forms *per se*. This approach has the advantage that it can be documented and reproduced easily and is not subject to *ad hoc* human intervention and interpretation. It

[16] For some possible models see Schomaker *Java Demos*.

therefore stands in contrast to other approaches which require the user to segment the image into letters or to build up large databases of graphemes (Ciula §§26–30; Bulacu and Schomaker 2007 282). This approach has also been used with some success on both modern and medieval material (Srihari et al. 2002; Stokes 2007/8; but compare Bulacu and Schomaker 2006 and Bulacu and Schomaker 2007 283–4), and indeed it has been proposed that such an approach is better than considering morphology when comparing scribal hands. "Even more than in ductus and sense of form and proportion the idiosyncratic is to be found in the production of single strokes, in the behaviour of the pen as it turns a curve or a corner, in features defying verbal analysis but offering a limit beyond which sustained imitation, with any appearance of spontaneity, becomes exceedingly difficult" (Bishop 1961 9). One might reasonably ask if this is where "digital" palaeography (as opposed to computer-aided) may be at its best: not by examining letter-forms which can be imitated and which vary anyway depending on the script being written, but instead by looking at the minutiae of strokes over a relatively large sample and which might reflect the individual writer whatever the script.

This raises a further question: even if these digital methods can be used to identify writers, what about the two other questions asked of palaeographers, namely where and when the sample was written? The techniques have been tested for this sort of use (Ciula), but they can only associate similar samples, and it is then up to the palaeographer to interpret the results. Indeed, it has been suggested that morphology, the forms of letters, is a more appropriate criterion for establishing chronological and geographical styles (Derolez 6–7; compare also Stokes 2005 but note Ciula §11), and it may well be that this sort of study requires both evidence that is less amenable to computerisation but much more to human interpretation than is the case for handwriting identification. Whether this is so requires investigation, and even if it is then one might hope that some of the methodological lessons learned here can be applied to these other problems as well.

Even the most objective method still necessarily involves interpretation, and this holds as much for the hard sciences as for the humanities. Palaeography, like every other field, therefore cannot ever be purely objective. However, the more we can articulate our methods and our results, the more we can debate our different interpretations, the more we can aid communication and interpretation and analysis, and the more quantitative and new evidence we can bring to the discussion, the stronger our conclusions will be.

Bibliography

Arts and Humanities Data Service (AHDS). AHRC Award Winners with Technical Components. 2006. <http://ahds.ac.uk/collections/ahrc-research-grant-winners.htm>.

Aussems, Mark. "*Christine de Pizan and the Scribal Fingerprint: A Quantitative Approach to Manuscript Studies.*" MA Thesis. Utrecht University, 2006. <http://igitur-archive.library.uu.nl/student-theses/2006-0908-200407/UUindex.html>.

Beneš, Carrie E. "The Appearance and Spread of the E-Cedilla." *Manuscripta* 43/44 (1999–2000): 1–43.

Benskin, Michael and Margaret Laing. "Translations and Mischsprachen in Middle English manuscripts." *So Meny People Longages and Tonges: Philological Essays in Scots and Medieval English Presented to Angus McIntosh.* Ed. Michael Benskin and Michael L. Samuels. Edinburgh: Middle English Dialect Project, 1981. 55–106.

Birsan, Dorian. "On Plug-ins and Extensible Architectures." *Queue (ACM)* 3:2 (2005): 40–6. <http://queue.acm.org/detail.cfm?id=1053345>.

Bischoff, Bernhard. *Latin Palaeography: Antiquity and the Middle Ages.* Trans. D. O Cróinín and D. Ganz. Cambridge: Cambridge University Press, 1990.

Bishop, T. Alan. *Scriptores Regis.* Oxford: Clarendon Press, 1961.

Bishop, T. Alan. *English Caroline Minuscule.* Oxford: Clarendon Press, 1971.

Bradley, John. "Collaborative Tool-Building with Pliny: A Progress Report." *Digital Humanities 2008: Book of Abstracts.* Ed. Lisa Lena Opas-Hänninen, Mikko Jokelainen, Ilkka Juuso and Tapio Seppänen. Oulu: University of Oulu, 2008. 65–7.

Brown, T. Julian. "Latin Palaeography Since Traube." *Transactions of the Cambridge Bibliographical Society* 3 (1959–63): 361–81.

Brown, T. Julian. *A Palaeographer's View: Selected Writings.* Ed. J. M. Bately, M. P. Brown and J. Roberts. London: Harvey Miller Publishers, 1993.

Brown, Michelle P. *The Book of Cerne: Prayer, Patronage and Power in Ninth-Century England.* London: British Library, 1996.

Bulacu, Marius, Lambert Schomaker, and Louis Vuurpijl. "Writer-Identification using Edge-based Directional Features." *Proceedings of the Seventh International Conference on Document Analysis and Recognition* 2 (2003): 937–41.

Bulacu, Marius and Lambert Schomaker. "Combining Multiple Features for Text-Independent Writer Identification and Verification." *Proceedings of the 10th International Workshop on Frontiers in Handwriting Recognition (IWFHR)* (2006): 281–6.

Bulacu, Marius and Lambert Schomaker. "Automatic Handwriting Identification on Medieval Documents." *Proceedings of the 14th International Conference on Image Analysis and Processing (ICIAP 2007).* Ed. Rita Cucchiara. Los Alamos, CA: IEEE Computer Society, 2007. 279–84.

Burgers, Jan W. J. *De paleografie van de documentaire bronnen in Holland en Zeeland in de dertiende eeuw.* Louvain: Peeters, 1995.

Case, Mary and David Green. "Rights and Permissions in an Electronic Edition." *Electronic Textual Editing.* Ed. Lou Burnard, Katherine O'Brien O'Keeffe, and John Unsworth. New York: Modern Language Association of America, 2006. 565–88.

Center of Excellence for Document Analysis and Recognition (CEDAR). *Individuality in Handwriting: Demo Programs.* CEDAR, State University of NY, 2008. <http://www.cedar.buffalo.edu/NIJ/programs.html>.

Ciula, Arianna. "Digital Palaeography: Using the Digital Representation of Medieval Script to Support Palaeographic Analysis." *Digital Medievalist* 1 (2005). <http://www.digitalmedievalist.org/journal/1.1/ciula/>.

Costamagna, Giorgio, et al. "Commentare Bischoff." *Scrittura e Civiltà* 19 (1995) 325–48 and 20 (1996) 401–7.

Craig-McFeely, Julia and Alan Lock. *Digital Image Archive of Medieval Music: Digital Restoration Workbook.* Oxford: Oxford Select Specialist Catalogue Publications, 2006.

Craig-McFeely, Julia. "Digital Image Archive of Medieval Music: The Evolution of a Digital Resource." *Digital Medievalist* 3 (2007/8). <http://www.digitalmedievalist.org/journal/3/mcfeely/>.

Crick, Julia. "St Albans, Westminster, and Some Twelfth-Century Views of the Anglo-Saxon Past." *Anglo-Norman Studies* 25 (2002): 65–84.

Cubbin, Geoffrey P., ed. *MS D: A Semi-Diplomatic Edition. The Anglo-Saxon Chronicle: A Collaborative Edition*, vol. 6. Cambridge: D. S. Brewer, 1996.

Davis, Lisa F. "Towards an Automated System of Script Classification." *Manuscripta* 42 (1998): 193–201.

Davis, Tom. "The Practice of Handwriting Identification." *The Library (7th series)* 8 (2007): 251–76.

Derolez, Albert. *The Palaeography of Gothic Manuscript Books from the Twelfth to the Early Sixteenth Century.* Cambridge: Cambridge University Press, 2003.

Drout, Michael D. C. "Anglo-Saxon Studies: The State of the Field?" *The Heroic Age* 11 (2008). <http://www.mun.ca/mst/heroicage/issues/11/foruma.php>.

Dumville, David N. "On the Dating of Some Late Anglo-Saxon Liturgical Manuscripts." *Transactions of the Cambridge Bibliographical Society* 10 (1991–95): 40–57.

Dumville, David N. *English Caroline Script and Monastic History: Studies in Benedictinism, A.D. 950–1030.* Woodbridge: The Boydell Press, 1993.

Fotoscientifica. <http://www.fotoscientificarecord.com>.

Ganz, David. "'Editorial Palaeography': One Teacher's Suggestions." *Gazette du Livre Médiéval* 16 (1990): 17–20.

Gilissen, Léon. *L'expertise des écritures médiévales. Recherche d'une méthode avec application à un manuscrit du XIe siècle: le Lectionnaire de Lobbes, codex Bruxellensis 18018.* Ghent: Éditions scientifiques E. Story-Scientia, 1973.

Gneuss, Helmut. "The Origin of Standard Old English and Æthelwold's School at Winchester." *Anglo-Saxon England* 1 (1972): 63–83.

Godden, Malcolm, Rohini Jayatilaka and Rosalind Love, eds. *Boethius in Early Medieval Europe: Commentary on The Consolation of Philosophy from the 9th to the 11th* *centuries.* Oxford: Faculty of English, 2007. <http://www.english.ox.ac.uk/boethius/>.

Gosling, James and Henry McGilton. *The Java Language Environment: A White Paper.* Sun Microsystems, 1996. <http://java.sun.com/docs/white/langenv/>.

Gretsch, Mechthild. "Winchester Vocabulary and Standard Old English: The Vernacular in late Anglo-Saxon England." *Bulletin of the John Rylands Library* 83 (2001): 41–88.

Guimon, Timofey V. "The Writing of Annals in Eleventh-Century England: Palaeography and Textual History." *Writing and Texts in Anglo-Saxon England.* Ed. Alexander R. Rumble. Cambridge: D. S. Brewer, 2006. 137–145.

Gullick, Michael. "The Hand of Symeon of Durham: Further Observations on the Durham Martyrology Scribe." *Symeon of Durham: Historian of Durham and the North.* Ed. David Rollason. Stamford: Shaun Tyas, 1998. 14–31.

Gumbert, J. Peter. "A Proposal for a Cartesian Nomenclature." *Essays presented to G. I. Lieftinck, IV, Miniatures, Scripts, Collections.* Ed. J. Peter Gumbert and Max J. M. de Haan. Amsterdam: Van Gendt, 1976. 45–52.

Gumbert, J. Peter. "Commentare 'Commentare Bischoff'." *Scrittura e Civiltà* 22 (1998): 397–404.

Hirtle, Peter. "Editorial." *D-Lib Magazine* 6.4 (2000). <http://www.dlib.org/dlib/april00/04editorial.html>.

Holmes, Martin. *The UVic Image Markup Tool Project.* <http://tapor.uvic.ca/~mholmes/image_markup/>.

Howorth, Henry H. "The Anglo-Saxon Chronicle, its Origin and History." *Archaeological Journal* 69 (1912): 312–370.

International Dunhuang Project (IDP): Technical Resources. London: British Library. <http://idp.bl.uk/pages/technical_resources.a4d>.

Intute. *Identification of Medieval Scribal Handwriting.* 2004. <http://www.intute.ac.uk/artsandhumanities/cgi-bin/fullrecord.pl?handle=humbul12875>.

JAI-Core Project Home Page. <https://jai-core.dev.java.net>.

Kam, Moshe, Joseph Wetstein, and Robert Conn. "Proficiency of Professional Document Examiners in Writer Identification." *Journal of Forensic Sciences* 39 (1994): 5–14.

Kam, Moshe, Gabriel Fielding, and Robert Conn. "Writer Identification by Professional Document Examiners." *Journal of Forensic Sciences* 42 (1997): 778–86.

Keller, Wolfgang. *Die litterarischen bestrebungen von Worcester in angelsächsischer zeit.* Strasbourg: K. J. Trübner, 1900.

Ker, Neil R. *Catalogue of Manuscripts containing Anglo-Saxon.* Oxford: Clarendon Press, 1957.

Kiernan, Kevin S. *Electronic Beowulf.* CD-ROM. London: The British Library, 1999.

Kingsbury, Nick. "The Dual-Tree Complex Wavelet Transform: A New Efficient Tool for Image Restoration and Enhancement." *Proc. European Signal Processing Conference, EUSIPCO 98, Rhodes* (1998): 319–22.

Lieftinck, Gerard I. "Pour une nomenclature de l'écriture livresque de la période dite gothique." *Nomenclature des écritures livresques du IXe au XVIe siècle: Premier colloque international de paléographie latine, Paris, 28–30 avril 1953.* Ed. Bernhard Bischoff, Gerard I. Lieftinck, and Giulio Battelli. Paris: Centre National de la Recherche Scientifique, 1954. 15–34.

Lieftinck, Gerard I., ed. *Manuscrits datés conservés dans les Pays-Bas: catalogue paléographique des manuscrits en écriture latine portant des indications de date. Tom. 1, Les manuscrits d'origine étrangère: 816–c.1550.* Amsterdam: North-Holland Publishing Co., 1964.

Lieftinck, Gerard I. and J. Peter Gumbert, eds. *Manuscrits datés conservés dans les Pays-Bas: Catalogue paléographique des manuscrits en écriture latine portant des indications de date, 2: Les manuscrits d'origine néerlandaise (XIVe-XVIe siècles) et supplément au tome premier.* Leiden: Brill, 1988.

Liptak, Adam. "Prosecutors Hope New Study of Handwriting Analysis will Silence Skeptics." *New York Times.* 26 May 2002: A14.

Maarse, Frans J. and Arnold J. W . M. Thomassen. "Produced and Perceived Writing Slant: Difference between Up and Down Strokes." *Acta Psychologica* 54 (1983): 131–47.

Mallon, Jean. *Paléographie romaine.* Madrid: Consejo Superior de Investigaciones Científicas, Instituto Antonio de Nebrija, de Filología, 1952.

McGillivray, Murray. "Statistical Analysis of Digital Paleographic Data: What Can it Tell Us?" *Computing in the Humanities Working Papers* A.33 (2005).
<http://www.chass.utoronto.ca/epc/chwp/Casta02/McGillivray_casta02.htm>.

McIntosh, Angus. "Towards an Inventory of Middle English Scribes." *Neuphilologische Mitteilungen* 75 (1974): 602–24.

McIntosh, Angus. "Scribal Profiles from Middle English Texts." *Neuphilologische Mitteilungen* 76 (1975): 218–35.

McIntosh, Angus, Michael L. Samuels, and Michael Benskin, eds. *A Linguistic Atlas of Late Mediaeval English.* Aberdeen: Aberdeen University Press, 1986.

McIntosh, Angus. "A New Approach to Middle English Dialectology." *Middle English Dialectology.* Ed. Margaret Laing. Aberdeen: Aberdeen University Press, 1989a. 22–31.

McIntosh, Angus. "Word Geography in the Lexicography of Middle English." *Middle English Dialectology.* Ed. Margaret Laing. Aberdeen: Aberdeen University Press, 1989b. 86–97.

MIX: NISO Metadata for Images in XML Schema. 2008.
<http://www.loc.gov/standards/mix/>.

METS: Metadata Encoding and Transmission Standard Official Website. 2009. <http://www.loc.gov/standards/mets/>.

Michel, Lothar. *Gerichtliche Schriftvergleichung: eine Einführung in Grundlagen, Methoden und Praxis.* Berlin: De Gruyter, 1982.

Muir, Bernard J. "A Preliminary Report on a New Edition of the Exeter Book." *Scriptorium* 43 (1989): 273–88.

National Information Standards Organization (NISO). *ANSI/NISO Z39.87 Data Dictionary: Technical Metadata for Digital Still Images.* Bethesda, MD: NISO Press, 2006. <http://www.niso.org/kst/reports/standards/>.

O'Donnell, Daniel P. "The Doomsday Machine, or, 'If you Build it, Will they Still Come Ten Years from Now?': What Medievalists Working in Digital Media Can Do to Ensure the Longevity of their Research." *The Heroic Age* 7 (2004). <http://www.mun.ca/mst/heroicage/issues/7/ecolumn.html>.

O'Donnell, Daniel P. "Disciplinary Impact and Technological Obsolescence in Digital Medieval Studies." *A Companion to Digital Literary Studies.* Ed. Ray Siemens and Susan Schreibman. Oxford: Blackwell, 2007. 65–81.

Padfield, Tim. *Copyright for Archivists and Records Managers.* London: Facet Publishing, 2007.

Parkes, Malcolm B. "Latin Autograph Manuscripts: Orthography and Punctuation." *Gli autografi medievali: problemi paleografici e filologici : atti del convegno di studio della Fondazione Ezio Franceschini, Erice, 25 settembre-2 ottobre 1990.* Ed. Paolo Chiesa and Lucia Pinelli. Spoleto: Centro italiano di studi sull'alto Medioevo, 1994. 23–36.

Parkes, Malcolm B. *Their Hands Before Our Eyes: A Closer Look at Scribes.* Aldershot: Ashgate, 2008.

Pierazzo, Elena. "Editorial Teamwork in a Digital Environment: The Edition of the Correspondence of Giacomo Puccini." *Jahrbuch für Computerphilologie* 10 (2008). <http://computerphilologie.tu-darmstadt.de/jg08/pierazzo.html>.

Pitti, Daniel V. "Designing Sustainable Projects and Publications." *A Companion to Digital Humanities.* Ed. Susan Schreibman, Ray Siemens and John Unsworth. Oxford: Blackwell, 2004. 471–87.

Plummer, C., ed. *Two of the Saxon Chronicles Parallel, with Supplementary.* Extracts from the Other Edition, 2 vols. Clarendon Press, Oxford, 1892–99. Revised, on the basis of an edition by John Earle.

Pratesi, Alessandro. "Commentare Bischoff: un secondo intervento." *Scrittura e Civiltà* 22 (1998): 405–8.

Rahmat-Samii, Yahya and Eric Michielssen, eds. *Electromagnetic Optimization by Genetic Algorithms.* New York: J. Wiley, 1999.

Rinascimento Virtuale. Digitale Palimpsestforschung. Rediscovering written records of a hidden European cultural heritage. <http://www1.uni-hamburg.de/RV/>.

Robinson, Pamela R. *Catalogue of Dated and Datable Manuscripts, c. 737–1600, in Cambridge Libraries.* Cambridge: D. S. Brewer, 1988.

Rumble, Alexander R. "Using Anglo-Saxon Manuscripts." *Anglo-Saxon Manuscripts: Basic Readings.* Ed. Mary P. Richards. New York: Garland, 1994. 3–24.

Rumble, Alexander R. "Palaeography, Scribal Identification and the Study of Manuscript Characteristics." *Care and Conservation of Manuscripts 8: Proceedings of the 8th International Seminar.* Ed. Gillian Fellows-Jensen and Peter Springborg. Copenhagen: Museum Tusculanum Press, 2005. 217–28.

Rumble, Alexander R. "The Study of Anglo-Saxon Manuscripts, Collections and Scribes: in the Footsteps of Wanley and Ker." *Writing and Texts in Anglo-Saxon England.* Ed. Alexander R. Rumble. Cambridge: D. S. Brewer, 2006. 1–17.

Russell, Paul, Sharan Arbuthnot, and Pádraic Moran. *Early Irish Glossaries Project.* Cambridge: Dept. ASNC, 2006. <http://www.asnc.cam.ac.uk/irishglossaries/>.

Schomaker, Lambert. "Advances in Writer Identification and Verification." *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR 2007), 26 September, Curitiba, Brazil* 2 (2007): 769–73.

Schomaker, Lambert. *Java Demos for Handwriting Recognizers.* University of Groningen. <http://www.ai.rug.nl/~lambert/recog/java-demos.html>.

Scragg, D. G., Alexander R. Rumble, Kathryn Powell, Susan D. Thompson, and Joana Soliva. *MANCASS C11 Database Project.* Manchester: Manchester Centre for Anglo-Saxon Studies, 2005. <http://www.arts.manchester.ac.uk/mancass/C11database/>.

Sculley, D. and Bradley M. Pasanek. "Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities." *Literary and Linguistic Computing* 23 (2008): 409–24.

Shippey, Tom. "Response to Three Papers on 'Philology: Whence and Whither?' Given by Drs Utz, Macgillivray, and Zolkowski, at Kalamazoo, 4th May 2002." *The Heroic Age* 11 (2008). <http://www.mun.ca/mst/heroicage/issues/11/foruma.php>.

Sirat, Colette. *Writing as Handwork: A History of Handwriting in Mediterranean and Western Culture.* Ed. Lenn Schramm. Turnhout: Brepols, 2006.

Srihari, Sargur N. *Handwriting Identification: Research to Study Validity of Individuality of Handwriting and Develop Computer-assisted Procedures for Comparing Handwriting.* New York: Center of Excellence for Document Analysis and Recognition, 2001.

Srihari, Sargur N., Sung-Hyuk Cha, Hina Arora, and Sangjik Lee. "Individuality of Handwriting." *Journal of Forensic Sciences* 47 (2002): 856–72.

Srihari, Sargur N., Chen Huang, and Harish Srinivasan. "On the Discriminability of the Handwriting of Twins." *Journal of Forensic Sciences* 53 (2008): 430–46.

Stokes, Peter A. "Shoots and Vines: Some Models for the Ascenders and Descenders of English Vernacular Minuscule." *Quaestio Insularis: Selected Proceedings of the Cambridge Colloquium in Anglo-Saxon, Norse, and Celtic* 5 (2004): 98–109.

Stokes, Peter A. "*English Vernacular Script, ca 990 – ca 1035*." PhD Thesis. Cambridge University, 2005.

Stokes, Peter A. "Palaeography and Image Processing: Some Solutions and Problems." *Digital Medievalist* 3 (2007/8). <http://www.digitalmedievalist.org/journal/3/stokes/>.

Stokes, Peter A. "Recovering Anglo-Saxon Erasures: Some Questions, Tools and Techniques." *Palimpsests and the Literary Imagination of Medieval England*. Ed. Raeleen Chai-Elsholz, Tatjana Silec and Leo Carruthers. New York: Palgrave, forthcoming 2009.

Stokes, Peter A. "Palaeography and the 'Virtual Library'." *Digitizing Medieval and Early Modern Material Culture*. Ed. Brent Nelson and Melissa Terras. Forthcoming.

TEI Consortium, The. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Ed. Lou Burnard and Syd Bauman. Oxford: Humanities Computing Unit, University of Oxford, 2009.

Traube, Ludwig. *Nomina sacra: Versuch einer Geschichte der christlichen Kürzung*. Munich: C. H. Beck, 1907.

United States v. Prime: Order Regarding Defendant's Motion In Limine. US Supreme Court, 9th Circuit, 2002. 220 F. Supp. 2d 1203 (W. D. Wash. 2002).

United States v. Janet L. Thornton. 2003. Case No. 02-M-9150-01, decided January 24, 2003.

Watson, Andrew G. *Catalogue of Dated and Datable Manuscripts, c. 700–1600, in the Department of Manuscripts, The British Library*. London: British Library, 1979.

Wright, Cyril E. *English Vernacular Hands from the Twelfth to the Fifteenth Centuries*. Oxford: Clarendon Press, 1960.

Zupitza, Julius. *Beowulf: Reproduced in Facsimile from the Unique Manuscript British Museum MS. Cotton Vitellius A.xv*. London: Oxford University Press, 1959.

Appendizes

Appendices

Kurzbiographien – Biographical Notes

Bernhard Assmann studierte an der Universität zu Köln Informationsverarbeitung, Mittlere und Neuere Geschichte und Historische Hilfswissenschaften. Danach betreute er das Digitalisierungsprojekt »Die Werke Friedrichs des Großen« an der Universitätsbibliothek Trier. Im Moment ist er beim Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen in Köln beschäftigt.

Mark Aussems is a PhD student at the University of Edinburgh. His doctoral thesis focuses on scribal identification in the fifty supervised presentation manuscripts of the works of Christine de Pizan. His research interests further include digital and quantitative palaeography and codicology. Aussems is also research associate to the Arts and Humanities Research Council (AHRC) research project »Christine de Pizan: the Making of the Queen's Manuscript«, as well as co-editor of the Dutch academic journal »Madoc. Tijdschrift over de Middeleeuwen«.

Francesco Bernardi si è laureato a Venezia in codicologia. Dal 2003 referente scientifico e tecnico di Nuova Biblioteca Manoscritta, Bernardi cataloga i manoscritti araldici della Biblioteca Correr, Venezia.

Axel Brink received an MSc degree in computer science from the University of Groningen in 2004 and completed one year of communication and information sciences in 2005. Since 2005, he has been with the Institute of Artificial Intelligence and Cognitive Engineering (ALICE) at the University of Groningen, pursuing a PhD degree. His research focuses on automatic writer identification for forensic and historical handwriting. His scientific interests include artificial intelligence, pattern recognition and interaction design.

Antonio Cartelli is a researcher in Didactics and Special Pedagogy. He manages the Laboratory for Teaching-Learning Technologies in the Department of Human and Social Sciences and the Centre for ICT and on line teaching in the Faculty of Humanities at the University of Cassino. He has published papers and books concerning themes like misconceptions and mental schemes in scientific disciplines and especially digital literacy, Information Systems and Web Technologies for research and teaching. He is the editor, in collaboration with Marco Palma, of the

Encyclopedia of Information Communication Technology, published in 2008 by Information Science Reference, USA.

Hugh Cayless holds a PhD in Classics and a Master of Science in Information Science, both from University of North Carolina (UNC), Chapel Hill. He is presently involved in creating a portal for papyrological research for the New York University Digital Library and the Institute for the Study of the Ancient World, under a grant from the National Endowment for the Humanities. He also serves as a co-investigator on the img2xml project at UNC.

Arianna Ciula has been a research associate at the Centre for Computing in the Humanities, King's College London from 2003 to 2008. She currently works as Science Officer (Humanities) at the European Science Foundation. Her research interests focus on the modelling of scholarly digital resources related to primary sources. She lectured on humanities computing, published on digital palaeography and digital philology, organised conferences in digital humanities, and is elected member of the Association for Literary and Linguistic Computing Executive Committee, Digital Medievalist Board, Text Encoding Initiative Council.

Andrea Daltri is an archivist at the University of Bologna. He has published essays about the history of Cesena in the 19th century. Since 2002, he has been the curator of the database project of the Open Catalogue of the Malatestiana Manuscripts and has overseen its implementation.

Daniel Deckers ist wissenschaftlicher Mitarbeiter und Koordinator des Teuchos-Zentrums am Institut für Griechische und Lateinische Philologie der Universität Hamburg. Seine Forschungsschwerpunkte sind Antikes Schriftwesen, multispektrale Handschriftendigitalisierung, gräzistische Palimpsestforschung, digitale Klassische Philologie und Editorik.

Paolo Eleuteri, professore di codicologia all'università di Venezia. Interessi scientifici principali: scritture greche, catalogazione dei manoscritti, storia e ricezione dei testi.

Paola Errani has been curator of manuscripts at the Malatestiana Library since 1999. She published a printed catalogue of the dated manuscripts owned by the Malatestiana and a short history of the foundation of the public library in Cesena. She is responsible for website management and administration for the Open Catalogue of the Malatestiana Manuscripts.

Franz Fischer ist als Mitarbeiter des St Patrick's Confessio Hypertext Stack Project an der Royal Irish Academy in Dublin tätig. Auf das Studium der Fächer Latein und Geschichte in Köln und Rom folgte eine Dissertation in Mittellateinischer Philologie mit der digitalen Edition der Summa de officiis ecclesiasticis Wilhelms von Auxerre. Daneben war er am Thomas-Institut und am Zentrum für Mittelalterstudien der Universität zu Köln sowie an der Ruhr Universität Bochum beschäftigt.

Christiane Fritze studierte Romanistik, Slavistik und Bibliotheks- und Informationswissenschaft in Leipzig, Rennes und Berlin. Seit 2001 ist sie als wissenschaftliche Mitarbeiterin der Berlin-Brandenburgischen Akademie der Wissenschaften – beim Digitalem Wörterbuch der deutschen Sprache, bei Telota (The electronic life of the academy) und seit 2007 als Projektkoordinatorin des Deutschen Textarchivs (DTA) – mit verschiedensten Aspekten der Digitalisierung und Transformation geisteswissenschaftlicher Inhalte konfrontiert und führt darüber hinaus Lehrveranstaltungen im Bereich der Digital Humanities durch.

Daniele Fusi holds a PhD in classical philology and teaches digital humanities at La Sapienza, University of Rome. He develops computer-based academic and commercial projects, and runs a small business IT company mainly involved with solutions specialized in linguistical and philological fields. His main interests are classical languages and metrics, studied with the aid of computer-based expert systems, epigraphical and literary editions with highly specialized and varied contents, digital lexicography, and software development.

Maria Gurrado is a researcher at the Institut de recherche et d'histoire des textes (IRHT), Centre national de la recherche scientifique (CNRS), Paris, in the palaeographical section. She is currently engaged in the GRAPHEM (Grapheme based Retrieval and Analysis for PaleograpHic Expertise of Middle Age manuscripts) projet. Current interests include Latin palaeography, quantitative palaeography, and cognitive neuro-science with regard to handwriting.

Wernfried Hofmeister lehrt im Bereich der Germanistischen Mediävistik an der Karl-Franzens-Universität Graz. Zu seinen Forschungsbereichen zählen neben der theoretischen wie praktischen Editionswissenschaft und der mentalitätszentrierten Untersuchung mittelalterlicher Literatur auch die diachrone Phraseologie und Schriftforschung. Er ist mehrfacher Projektleiter, Reihenherausgeber, (Vorstands-) Mitglied in internationalen Fachvereinigungen und erhielt renommierte Förderungspreise zuerkannt. 2002 nahm er eine Gastprofessur in Minnesota wahr.

Andrea Hofmeister-Winter forscht und lehrt als Assistentin am Institut für Germanistik der Karl-Franzens-Universität Graz. Ihre wissenschaftlichen Schwerpunkte liegen im Bereich der theoretischen und praktischen Editorik sowie der historischen Graphetik. Sie ist maßgeblich am Grazer Projekt »Datenbank zur Authentifizierung mittelalterlicher Schreiberhände« (DAmalS) beteiligt, dessen Konzept von ihr mit entwickelt wurde.

Pamela Kalning ist wissenschaftliche Angestellte an der Universitätsbibliothek Heidelberg. Sie arbeitet im Rahmen eines DFG-Projektes an der Katalogisierung der deutschsprachigen Palatinahandschriften. Ihr Forschungsschwerpunkt lag bisher auf der didaktischen Literatur des späten Mittelalters, insbesondere veröffentlichte sie über Kriegslehren (Seffner, Rothe, Wittenwiler) und Schachzabelbücher. Sie ist Mitherausgeberin der 2009 erschienenen, mit einer Übersetzung und einem ausführlichen Kommentar versehenen Edition von Johannes Rothes »Ritterspiegel«.

Silke Kamp ist Promotionsstudentin an der Technischen Universität Berlin und arbeitet zum Thema »Hugenotten in Potsdam 1685-1809«. Weitere Forschungsschwerpunkte sind Arbeit und Magie in der Frühen Neuzeit sowie Frauen in Brandenburg-Preußen. Sie war im Wintersemester 2008/2009 Lehrbeauftragte am Historischen Institut der Universität Potsdam.

John Keating is the Associate Director of An Foras Feasa, The Institute for Research in Irish Historical and Cultural Traditions. His research interests include humanities computing, historical manuscript encoding, hyperspectral segmentation, software engineering and new methods of information transfer, for example, the Strangers to Citizens virtual exhibition currently available in the National Library of Ireland. He is one of the principle investigators of An Foras Feasa's Higher Education Authority (HEA) funded grant »Humanities Serving Irish Society: Humanities, Technology, Innovation« (HSIS).

Adolf Knoll is director for science, research, and international cooperation of the National Library of the Czech Republic. As a deputy director, Knoll has been responsible for many digitization-related R&D projects. He is a member of the High Level Expert Group on Digital Libraries at the European Commission, of the UN-

ESCO Memory of the World Sub-Committee on Technology, and of The European Library Management Committee.

Lutz Koch ist wissenschaftlicher Mitarbeiter am Institut für Griechische und Lateinische Philologie der Freien Universität Berlin und koordiniert die Arbeiten am Aristoteles-Archiv im Rahmen des Teuchos-Zentrums (Universität Hamburg). Seine Forschungsschwerpunkte sind griechische Paläographie, Textüberlieferung, Editorik, Geschichte der antiken Philosophie und Wissenschaft, Aristoteles, Ptolemaios, Philologiegeschichte sowie digitale Philologie.

Bernard J. Muir is Professor of Medieval Studies at the University of Melbourne. He is Director of the digital publishing company Evellum and founder and producer of Bodleian Digital Texts. Two current projects of Evellum focus on the workings of the medieval scriptorium; with Andrew Turner he has just completed a digital facsimile edition of Terence's Six Latin Comedies and produced The Vernon Manuscript for the Bodleian Digital Texts series.

Marco Palma is professor of Latin palaeography at the University of Cassino. He has published some catalogues of dated manuscripts in Italian libraries. He defined with Antonio Cartelli the Open Catalogue of Manuscripts information system, which has been adopted by the staff of the Malatestiana Library.

Malte Rehbein is a Marie Curie Research Fellow at the National University of Ireland, Galway, and member of the Transfer of Expertise in Technologies of Editing (TEXTE) programme. He is a graduate in history and mathematics from the University of Göttingen, Germany and spent some years in industry as a software developer, project manager and consultant.

Patrick Sahle hat Geschichte, Philosophie und Politik in Köln und in Rom studiert. In den letzten zehn Jahren hat er für und mit einer Reihe verschiedenster universitärer Lehrstühle (Geschichte, Informatik), Bibliotheken, Archive und Museen gearbeitet. Zu den Arbeitsschwerpunkten gehörten immer wieder die Digitalisierung, Erschließung und Edition historischer Dokumente. Dabei war er u.a. am Projekt Codices Electronici Ecclesiae Coloniensis (CEEC) beteiligt. Zur Zeit ist er Lecturer in Humanities IT an der Universität zu Köln, wo er auch seine Dissertation zum Thema »Digitale Editionsformen« im Mai 2009 eingereicht.

Torsten Schaßan ist Mitarbeiter an der Abteilung Handschriften, Inkunabeln und Sondersammlungen der Herzog August Bibliothek Wolfenbüttel. Er betreut dort die digitale Erschließung der historischen Bestände, insbesondere die Handschriftenkatalogisierung und digitale Editionen. Weiterhin ist er Mitarbeiter des Digitalisierungsprojektes »e-codices«. Er studierte in Köln Mittlere und Neuere Geschichte, Germanistik und Philosophie.

Patrick Shiel is a PhD student in the Department of Computer Science at National University of Ireland, Maynooth. His current research, funded by An Foras Feasa, The Institute for Research in Irish Historical and Cultural Traditions, includes the automated hyperspectral segmentation of historical documents. Shiel received a first-class degree in computer science and software engineering in 2007, and has worked as a software developer for the Irish Research Council for the Humanities and Social Sciences (IRCHSS) funded history project »The Associational Culture in Ireland«.

Christian Speer studierte von 1996 bis 2003 Mittelalterliche Geschichte, Alte Geschichte und Kunstgeschichte in Dresden und Rom. Von 2004 bis 2008 war er wissenschaftlicher Mitarbeiter am Max-Plank-Institut für Geschichte Göttingen. 2005/06 war er Gastwissenschaftler am Centro per gli studi storici italo-germanici in Trient und am Department of History der University of California Berkeley. Seit 2008 ist Christian Speer wissenschaftlicher Mitarbeiter in der Handschriftenabteilung der Thüringer Universitäts- und Landesbibliothek Jena. Sein Forschungsschwerpunkt sind Handschriften des 14.–16. Jh. zur Kirchen-, Frömmigkeits- und Sozialgeschichte in Mitteldeutschland.

Mark Stansbury is a lecturer in Classics at the National University of Ireland, Galway, where he teaches a palaeography module. His research interests focus on the manuscript culture of Late Antiquity and the Early Middle Ages. He has written on the commentary tradition and is co-translator of Servius' commentary on Book 4 of the Aeneid. He is also co-director of the Columbanus Life and Legacy project at the Moore Institute.

Timothy Stinson is an assistant professor of English at North Carolina State University and co-editor of The Siege of Jerusalem Electronic Archive. Recent research includes an article on the relationship of early English printing to manuscript culture published in the Yearbook of Langland Studies and a forthcoming piece in the

Papers of the Bibliographical Society of America that documents his development of techniques to extract and analyze the DNA found in medieval parchment.

Peter Stokes is Leverhulme Early Career Fellow at the Department of Anglo-Saxon Norse and Celtic in the University of Cambridge, where he is developing new quantitative and computer-based methods in palaeography. His other primary interests include the vernacular English scripts of the late-tenth through twelfth centuries, and he has also published on computing in lexicography, Anglo-Saxon charters and bounds, and early-modern book collectors, as well as developing software for digital humanities.

Georg Thallinger leitet am Institut für Informationssysteme der Joanneum Research den Bereich Digitale Medien. Er beschäftigt sich mit der Umsetzung innovativer Lösungen im Bereich Filmrestauration, inhaltsbasierte Indizierung und Suche von audiovisuellen Medien, Medienmonitoring und Sicherheit. Besonderes Augenmerk legt er bei seiner Arbeit darauf, Anwendungsgebiete in gänzlich anderen Bereichen aufzutun und mit Forschergruppen aus z.B. den Geisteswissenschaften zusammenzuarbeiten. Neben der inhaltlichen Arbeit ist als wesentlicher Schwerpunkt die Koordination großer, internationaler Projekte zu nennen.

Gilbert Tomasi, Génie-Physique diplôme (Institut National des Sciences Appliquées de Lyon), Dr. Ing. (Technische Universität Berlin), former researcher at the Technische Universität Berlin, development manager at Siemens AG in Munich (patents), product marketing director at Thomson-CSF for Germany and Austria, owner of the former German company »Ingenieurbüro Dr. Tomasi« that distributed high-tech products from French speaking countries, specialist for semiconductor technology and microelectronics; CEO of the French software company »B.I.T. Bureau Ingénieur Tomasi«.

Roland Tomasi studied applied mathematics at the Technische Universität München and studies mathematics at Ludwig-Maximilians-Universität München; owner of three patents in the field of OCR and pattern recognition, R&D manager and software engineer working for the company »B.I.T. Bureau Ingénieur Tomasi«, owner and author of all B.I.T. source-codes.

Zdeněk Uhlíř, coordinator of the Manuscriptorium digital library and deputy director of the historical and music collections of the National Library of the Czech Republic, has been responsible for national and international standardisation projects concerning electronic manuscript description and for integration of resources providing written cultural heritage. Uhlíř has published books and articles on manuscript studies, medieval hagiography and homiletics as well as scholarly editions of medieval texts.

Barbara Vanin, laurea in lettere e beni culturali, fino al 2008 responsabile dei manoscritti alla Biblioteca Correr, dottoranda in Scienze umanistiche all'università di Venezia. Coordinatrice di Nuova Biblioteca Manoscritta. Cataloga i manoscritti medievali in volgare della Biblioteca Correr.

Cristina Vertan ist wissenschaftliche Mitarbeiterin am Teuchos-Zentrum des Instituts für Griechische und Lateinische Philologie der Universität Hamburg, sowie Co-Leiterin der interdisziplinären Arbeitsgruppe »Computerphilologie«. Sie wurde im Fach Informatik an der Universität Bukarest promoviert. Von 2002-2003 war sie Humboldt-Stipendiatin an der Universität Hamburg. Zwischen 2003 und 2007 leitete sie mehrere durch die EU und national geförderte Projekte im Bereich Maschinelle Sprachverarbeitung. Ihre aktuellen Forschungsschwerpunkte sind Computergestützte Modellierung und Management von heterogenen und multilingualen Daten in den Geisteswissenschaften, maschinelle Übersetzung sowie multilinguale Informationsauffindung.

Georg Vogeler ist wissenschaftlicher Mitarbeiter an der Professur für Historische Grundwissenschaften und Medienkunde der Ludwig-Maximilians-Universität München. Nach seinem Studium der Geschichtlichen Hilfswissenschaften, der Sozial- und Wirtschaftsgeschichte und des öffentlichen Rechts beschäftigte er sich an den Universitäten München und Lecce mit spätmittelalterlichen Amtsbüchern, den Urkunden Kaiser Friedrichs II. und der Digitalen Diplomatik. Er ist seit 2008 Mitglied des Moderamen der Association Paléographique Internationale Culture Écriture Société (APICES).

Christina Wolf ist als Projektbearbeiterin im Landesarchiv Baden-Württemberg in Stuttgart tätig. Derzeit besteht ihre Hauptaufgabe in der Betreuung des europäischen Projekts »Europeana. Aufbau einer europäischen digitalen Bibliothek« und in der Unterstützung des Bundesratsbeauftragten für die Digitalisierung von Kulturgut. Zuvor koordinierte sie für den archivischen Bereich die EU-Projekte »MICHAEL Plus« und »Bernstein«. Darüber hinaus ist sie archivische Ansprechpartnerin für das BAM-Portal (Gemeinsames Portal zu Bibliotheken, Archiven und Museen).

Paolo Zanfini has been a librarian at the Malatestiana Library since 2007. Since 2002, he has been the curator of the website project of the Open Catalogue of the Malatestiana Manuscripts and has overseen its realization.

Karin Zimmermann ist nach dem Studium der Germanistik und der Evangelischen Theologie an der Universität Heidelberg seit 1996 Mitarbeiterin im Projekt zur Rekatalogisierung der Codices Palatini germanici. Seit Ende 2008 ist sie stellvertretende Leiterin der Abteilung Handschriften und Alte Drucke der Universitätsbibliothek Heidelberg. Schwerpunkte ihrer Arbeit liegen in der Handschriftenkunde, der Katalogisierung der Codices Palatini germanici und der Digitalisierung und Erschließung des historischen Altbestandes der Universitätsbibliothek Heidelberg.

Handschriften-Register – Index of manuscripts