White Paper Report

Report ID: 98416

Application Number: HJ5000110

Project Director: Dean Rehberger (rehberge@msu.edu)

Institution: Michigan State University

Reporting Period: 1/1/2010-3/31/2011

Report Due: 6/30/2011

Date Submitted: 7/11/2011

Digging into Image Data to Answer Authorship-Related Questions

Peter Ainsworth – Peter Bajcsy – Dean Rehberger[1]

In an enthralling novel about gold prospecting in mid-19th-century New Zealand, Rose Tremain describes the 'grim business' of fossicking:

> For there was no other way of getting at the gold than by digging and panning what was known as the 'wash-dirt' or the 'pay-dirt'. Fossicking for the color was a grimy business. For every tiny ounce of gold there was a huge, embarrassing pile of detritus.
>
> The weeks began to pass. There was a difference between finding nothing whatsoever and the 'almost nothing' that was a minute pinch of powdery color, painstakingly rinsed from the soil and collected in a rag. And most days Joseph found nothing. Then he would experience a fury he knew was out of all proportion to his circumstances and he would bitterly recall his father saying: 'Joseph, in every life, desire may be frustrated. Try to be more accommodating to the world when it crosses you.' He found it very hard to accommodate the maddening absence of what he knew *had to be there* – if only he could see it.[2]

The challenge that this particular research group has been addressing over the past two years has been similar in some ways, though our frustrations have had more to do with the sheer *plenty* rather than scarcity of the seams we've sought to quarry. What Joseph's operation and our own have in common is the unremitting quest for the *color*. In his case, the 'glister' of gold; in ours, the identification of those properties which data-mining alone might, we trust, bring fully into the light, albeit beneath the already solid strata laid by more traditional but no less rigorous modes of scholarship.[3]

Contents

[1] The present paper is signed by the three project PIs, but alongside each of them is a formidably talented team of colleagues; all of them have contributed in diverse ways to the processes and outcomes described below. See the acknowledgment section of this paper.
[2] Rose Tremain, *The Colour*, Chatto and Windus-Vintage (London-NY, 2003), p. 73.
[3] One normally builds *on* foundations of course, but the requirements of sustaining the data-mining metaphor need somehow to be maintained…

1 Responding to the Call for Proposals

The announcement barely three years ago by the international consortium formed by the NEH, NSF, JISC and SSHRC of a programme to promote the application to arts and humanities scholarship of the emergent methodologies and specificities associated with what is now widely badged as e-Science[4] was both welcome and timely. Data-mining has been around for more than a decade; particle physicists, bio-scientists and engineers have been using data grids for even longer. But the past five years have seen exponential growth in the still relatively new and exciting domain of e-Science for the Arts and Humanities and Social Sciences, and an equally rapid uptake of its novel approaches and perspectives by scholars eager to grasp the promise of being able to work 'outside the silo' with practitioners from disciplines they would normally never have encountered.[5] Interdisciplinarity can of course be a mere buzzword, passing fashion or temporary ripple spreading outwards from some (visionary) funding council. Genuine interdisciplinarity, in contrast, promotes and facilitates outcomes that really are 'more than the sum of their discrete parts'.
That the quadroon of funders of *Digging into Data* should have decided to launch a second funding round with new additional funders bespeaks a confidence that projects of early promise are beginning to bear fruit, and that this will lead in due course to even more productive endeavours.

The immediate challenge of this white paper is to offer a cogent and coherent record of what our particular research collective has attained during the 15-month

[4] We refer in particular to a keynote address by Professor Daniel E. Atkins of the University of Michigan at a JISC-hosted symposium: 'The International e-Science Movement: Status and Future' (Edinburgh, 1st July 2010).
[5] One can quickly gauge the progress made in these areas by exploring recent issues of Digital Humanities Quarterly; see in particular, for present purposes: Ainsworth, Peter, and Michael Meredith, 'e-Science for medievalists: options, challenges, solutions and opportunities', Digital Humanities Quarterly, vol 3 number 4 (2009), issue on e-Science for the Arts and Humanities (ed. S. Dunn and T. Blanke), 12 p.

funding span with the resources allocated to us, and to show how it has empowered us to use the new methodologies described below, the better to refine and build on those more traditional methodologies already familiar to us from the fields of art history and connoisseurship, manuscript studies (including codicology and palaeography), cartography and history, and material culture, folklife and quilt history studies.

2 The Unifying Research Question of the Consortium

Digging into Image Data to Answer Authorship-Related Questions (or DID-ARQ as we shall henceforward call ourselves in this paper) came together as a research consortium in large measure on account of the especially favourable e-Science conjunction afforded by three lively research operations each with access to its own large-scale image collection. A strong, evolving track record in innovative research and investigation, allied to an increasing curiosity about what digital images (in addition to, but also in contrast to the actual, original objects photographed) might bring to the process of adducing novel research questions, was complemented by the resources and expertise afforded, respectively, by the Image Spatial Data Analysis Group (ISDA) and the Institute for Computing in Humanities, Arts, and Social Science (I-CHASS) at the National Center for Supercomputing Applications (NCSA) at Urbana-Champaign, The University of Illinois at Urbana-Champaign's University Library, their distinguished Colleges of Liberal Arts and Sciences and Fine and Applied Arts with its Center of excellence in History, Art History and Geography, MATRIX: The Center for Humane Arts, Letters, & Social Sciences Online, Michigan State University Museum (MSUM), and the Quilt Index,, and the Humanities Research Institute at the University of Sheffield.[6] A further key element was access to major collections of maps, quilts and medieval manuscripts.

In this manner was woven together the three-stranded cord for our data-mining endeavours; engagement with these has been undergirded throughout by a common preoccupation with issues of 'authorship' – in the context of three varieties of artifact that are typically anonymous, but made by highly skilled individuals or teams of such. While the three image collections (manuscripts, maps and quilt photographs) represent a very broad spectrum of historical content, the research question of authorship (artists, scribes, cartographers, or quiltmakers) has unified the team across departments, institutions, and international boundaries.

3 Getting down to work

The span of time available to us between hearing that we had received an award and officially launching the project was just a few days. Preliminary planning meant that we were nevertheless ready to 'hit the ground running'. Even so, the process of establishing a three-way virtual research partnership using shared data collections (to be accessible and editable in principle by all partners) proved a considerable

[6] Image Spatial Data Analysis Group: http://isda.ncsa.uiuc.edu/; Institute for Computing in Humanities, Arts, and Social Science: http://ichass.illinois.edu/; National Center for Supercomputing Applications: http://ncsa.illinois.edu; University of Illinois Libraries: http://www.library.illinois.edu/; MATRIX: http://www.matrix.msu.edu/; Quilt Index: http://www.quiltindex.org; MSU Museum: http://www.museum.msu.edu; and Humanities Research Institute: http://www.sheffield.ac.uk/hri/

challenge. Now that DID2 has been announced, the consortium hopes that a Technical Paper joint-authored by the DID-ARQ team may serve as a useful resource and benchmark for setting up similar projects in the not-too-distant future. Covering such issues as copyright clearance, communications infrastructure for coordination and knowledge sharing, intellectual property ownership, collaborative data and software sharing, attribution of credit in published results, as well as approaches to overcoming academic cultural differences – all within an extremely tight timeline, the writing of this paper[7] afforded the group an opportunity to reflect on the principal technical and intellectual challenges it faced, and to begin roughing out some of the solutions and outcomes emerging from the research conducted to date. A second opportunity was afforded by the announcement of the *Digging into Data* conference for Washington DC in June 2011: two years from the inception of our collective labours we now provide below a summary of the same research, from the standpoint this time of the Arts and Humanities researchers. Key outcomes, conclusions and findings are beginning to sharpen in focus, and some of these are presented in brief, prior to the publication of more formal articles in the appropriate journals.

3.1 Working with Manuscripts: the Froissart corpus

In the spring of 2011 an exhibition at the Hôtel National des Invalides in Paris curated by Peter Ainsworth in partnership with the Royal Armouries UK and the national French Musée de l'Armée portrayed aspects of the literary and military culture of the Hundred Years' War via items of weaponry from the period displayed against the backdrop provided by large-scale, high-resolution photographs of miniatures from Jehan Froissart's colorful *Chronicles*. Manuscript culture also featured in the guise of pens, brushes, pigments and other implements used by the scribes and artists responsible for the miniatures found in many manuscripts of the *Chronicles*; these resources were loaned by the remarkable *Scriptorial* museum in Avranches, Normandy. The centrepiece of the exhibition itself was a display in two cases facing one another across a darkened room, of two pairs of twin manuscripts. Besançon Municipal Library mss 864 and 865 (comprising respectively Book I and Books II-III of the *Chronicles*) thus found themselves just yards away from Paris, Bibliothèque nationale de France fonds français mss 2663 and 2664 (Book I and Book II of the *Chronicles*). This was exciting for Froissart scholars, aware that the four volumes had originally been copied and illustrated in Paris around 1412-1418 under the direction of a bookseller by the name of Pierre de Liffol.[8] The four volumes were displayed open at a fixed point, but their entire contents could be explored in virtual format via interactive touchscreens nearby, using the *Kiosque* software specially developed for the purpose at Sheffield's Humanities Research Institute.[9]

The virtual versions of the Besançon and Paris manuscripts were part of a corpus consisting of more than 6,000 high-resolution image files captured photographically from ten digitised manuscript volumes (2TB of data). The original manuscripts were all produced during the first quarter of the fifteenth century, many

---

[7] 'Digging into Data Using New Collaborative Infrastructures Supporting Humanities-based Computer Science Research', under consideration for *First Monday.*

[8] 'Les Chroniques de Froissart', *Art de l'enluminure*, nº 31, déc. 2009-fév. 2010, Editions Faton (Paris, 2009); G. Croenen, 'Le libraire Pierre de Liffol et la production de manuscrits illustrés des *Chroniques* de Jean Froissart à Paris au début du XVe siècle', pp. 14-23.

[9] http://www.shef.ac.uk/hri/projects/projectpages/kiosque/overview.html [accessed14 April 2011]

under the supervision of Pierre de Liffol who seems to have glimpsed a market opportunity for the production of luxury copies of Froissart's Middle French *Chronicles*. The *Chronicles* remain one of the most important prose accounts of the conflict composed in medieval France; they remain a key source for study of the Hundred Years' War between France and England and their respective allies. Their content forms the basis of a major new electronic resource, the *Online Froissart*.[10]

The DID-ARQ project has engaged above all with the virtual *Chronicles*, considered no longer simply as more flexible surrogates for the originals locked in their closed display cases or hidden away in their several research libraries, but rather as an additional source of data. Relationships carefully fostered with conservators and librarians, and the skills of a talented specialist photographer, led to the capture – under almost identical conditions – of no less than ten complete facsimiles (viewable via the *Online Froissart*), seven of them at 500 dpi. Development of fit-for-purpose manuscript viewing and manipulation software followed, *Virtual Vellum*;[11] finally copyright clearance was secured for shared use across the DID-ARQ consortium of all but two of the virtual manuscripts.

Colleagues at Illinois's College of Fine and Applied Arts, in partnership with NCSA, have adopted as their particular focus the virtual manuscripts' decorative and illustrative content, concentrating in particular on the twin manuscripts displayed in Paris last spring. Art historians working on early fifteenth-century iconography, artists and artistic schools in Paris long ago identified the hand responsible for the primary decoration of Besançon BM, ms. 864 and Paris, BnF f. fr. ms. 2664 as being that of a disciple of the Rohan Master. The artist responsible for the miniatures of Besançon BM, ms. 865 and Paris, BnF f. fr. Ms. 2663, on the other hand, was for many years thought to be a mediocre follower of the Master of the Berry Apocalypse, so called after a copy of the Apocalypse illustrated for the Duke of Berry and housed today at the Pierpont Morgan Library in New York (as ms. M.133)[12]. More recent scholarship, and in particular that conducted by Inès Villela-Petit, has argued that the Boethius Master deserves to be clearly distinguished from the Master of the Berry Apocalypse.[13] The artist illustrations to Besançon BM, ms. 864, meanwhile, are to be attributed to a forerunner of the Rohan Master, named the Giac Master on account of his having illustrated a book of Hours for Jeanne du Peschin, dame de Giac.[14]

The DID-ARQ humanists' starting point, therefore, was four manuscript volumes whose miniatures, initial letters and decorative borders were entrusted to two artists' workshops: those of the GIAC and BOETHIUS Masters. These artists were particularly favoured by bookseller De Liffol: their handiwork can be explored in several more of the virtual manuscripts forming our electronic corpus. Their artistry, and the often elegant penmanship of the scribes given the job of copying the texts, are an eloquent testimony to the remarkable activities of book trade artisans in Paris during the first quarter of the fifteenth century. The research questions Illinois began to explore included the following (expressed here in broad terms, with more detail provided below):

---

[10] http://hrionline.ac.uk/onlinefroissart [version 1.2; accessed 14 April 2011]
[11] http://www.shef.ac.uk/hri/projects/projectpages/virtualvellum.html [accessed 14 April 2011]
[12] This was for many years the view imposed by the eminent scholarship of Millard Meiss, *French Painting in the Time of Jean de Berry: The Limbourgs and their Contemporaries* (New York, 1974), chap. XI, p. 360 et sq.
[13] I. Villela-Petit, 'Deux visions de la *Cité de Dieu* : le Maître de Virgile et le Maître de Boèce', *Art de l'enluminure* N° 17, juin-juillet-août 2006, Editions Faton (Paris, 2006), pp. 2-19.
[14] I. Villela-Petit, 'Les Heures de Jeanne du Peschin, dame de Giac. Aux origines du Maître de Rohan', *Art de l'enluminure* N° 34, sept.-oct.-nov. 2010, Editions Faton (Paris, 2010), pp. 2-25.

- How does the application of computer algorithms to the analysis of portrayal of the human face in the manuscript miniatures help scholars to refine the parameters of discriminating features traditionally used by art connoisseurs for characterising the distinctive handiwork of individual artists such as the Giac and Boethius Masters?
- What do these computer techniques and their application reveal about the hands responsible for secondary decoration (e.g. initials or marginal decoration) of these manuscripts?
- To what extent could our new e-science techniques assist scholars in refining our knowledge of the human presence behind such broad-brush labels as 'Giac Master' or 'Boethius Master'? Do our procedures suggest the presence and activity of more than one individual active under these labels.

The Sheffield operation, meanwhile, concentrated on the copying process and its outcome: the writing or script constituting the text of the four volumes. Bookseller Pierre de Liffol produced these sumptuous illustrated books mainly for clients in the service of Charles VI of France, though the iconographical emphasis of the illustrations to at least one of them testifies to a client with pro-*English* sympathies. As for the text, this was copied from exemplars by at least two teams of scribes to whose workshops the unbound copies were sent by Pierre de Liffol[15].

Scholars tend to call these scribes 'A' or 'C', or even 'G', so anonymous are they for the most part[16]. Humanist scholars have none the less gradually built up an idea of particular scribal 'personalities' by singling out and describing the distinguishing features of their particular hands. Scholars have been able in this way to adduce significant characteristics on which to found their conclusions: particular ways of executing certain sequences of letters, such as the ligatures 'th', 'ch' or '-ent', say, or certain ways of writing the letters 'a' and 'r' (against the models furnished by contemporary bookhands taught to all apprentice scribes). The DID-ARQ digital archive comprises 10 virtual manuscripts each of which contains 300,000 words or more, all copied in a variety of one particular bookhand known to specialists as *littera cursive libraria*. DID-ARQ began testing some of the hypotheses adduced by palaeographers against this dataset, applying algorithms described in greater detail below. During the Royal Armouries exhibition in 2007-08 calligrapher Sara Mack offered live palaeography sessions during which she copied folios from the Stonyhurst manuscript of Froissart's *Chronicles*; her work provided the Sheffield DID-ARQ team with an enhanced understanding of the kinesis and *ductus* (see below) characterising the scribe responsible for that codex, which in turn informed the formulation of our research questions and overall approach. The key research question

---

15 The unbound quires were also sent out to the illustrators; their circulation and the piece work character of their gradual preparation prior to receiving the attentions of the bookbinder and -seller are key aspects of the book culture of this period, as evidenced in: *Patrons, Authors and Workshops. Books and Book Production in Paris around 1400*, G. Croenen and P. Ainsworth (ed.), Peeters, "Synthema" 4 (Louvain – Paris- Dudley MA, 2006). See also Anne D. Hedeman, *Translating the Past. Laurent de Premierfait and Boccaccio's* De casibus, J. Paul Getty Museum (Los Angeles, 2008).

16 There are of course notable exceptions; see for instance M.-H. Tesnière, 'Les manuscrits copiés par Raoul Tainguy : un aspect de la culture des grands officiers royaux au début du XV^{e} siècle', *Romania* 107 (1986), pp. 282-368. The transcriptions of the *Online Froissart* include marked-up indications wherever possible of shifts in hand from one copyist to another (see also *Online Froissart*, Apparatus, Codicological Descriptions).

(here, broadly formulated) underpinning our research theorization and practice in this domain was:

- How might one adduce pertinent e-Science methodologies for the interrogation of such a large-scale database, the better to explore, characterize and circumscribe several particular manifestations (individual scribal hands 'A', 'B', 'C'…) of an attested early 15th-century bookhand (*littera cursive libraria*)[17]?

3.2 Doing the Science: the Froissart corpus

3.2.1 Artist's Hands Contributing to Manuscripts

Using the Froissart manuscripts corpus provided by Sheffield (Department of French) and its library partners, specialists in Art History at Illinois supported by e-Science practitioner colleagues took as their objective two iconographic elements generally used by art historians as an index of Authorship. They started with two complementary quests: (i) for consistency of colors used for depicting the faces of queens, kings, and other figures within the illuminations, and (ii) for discriminating color space representations to separate two and more artistic hands. To the faces were applied algorithms used for measuring different or differential color spaces: RGB (CMY), HSV, YIQ, YUV, XYZ, LAB and LUV. Original RGB images from the database were subjected to color space conversion, producing an HSV image. Segmentation algorithms generated a Mask Image. Each of these two outcomes was then subjected to statistical analysis ($^{\mu}$H, $^{\mu}$S, $^{\mu}$V). A strong indication of the color palette characteristic of each particular artistic hand emerged from this analysis. Similar procedures applied to modern students' artwork generated equally sharp differentiations but the results have not been summarized in a publication yet.

Work has most recently focused on quantifying the color space distortions due to image manipulations (e.g., by looking at how the screen captures of the originals were taken, at the impact of changing file formats and of the introduction of file formats with lossy compression, as well as the passing of images into PowerPoint slides,etc.).

3.2.2 Scribal Hands

Traditional humanist scholarship by palaeographers suggested that it would be of interest and significance to compare letter and word clusters across our many hundreds of digitized folios of writing by a postulated scribe 'δ', for instance, using semi-automated definition of the perimeters of the letter and word shapes; this, it was postulated, should help to generate augmented – and more objective – evidence towards the assignment to a particular scribe of responsibility for a given section of a given manuscript, 'X'. Moreover once that scribe's written 'idiolect' had been so defined, it should become possible to search for his/her activity in other manuscripts, 'Y' or 'Z'. Scholars already have some professional evidence to suggest that this is happening across the codices of the 'Pierre de Liffol' corpus, but it was our

[17] See Jacques Stiennon, *Paléographie du Moyen Âge*, Armand Colin (Paris, 3rd edn, 1999), pp. 284-5; also Michelle P. Brown, *A Guide to Western Historical Scripts from Antiquity to 1600*, The British Museum (London, 1990).

expectation that finer and more accurate methodologies based on the virtual codices could confirm the hypothesis and account for it on the basis of more overtly scientific evidence. Potentially rich areas for electronic investigation included the *ductus* of the written text (the kinetic movement and direction of the hand and pen as they make their marks, their upward and downward strokes and curves) as realised by a particular scribe, the overall neatness of the particular folio, the (characteristic) recourse by the scribe to abbreviations, and their regular deployment of particular spelling patterns.

The Sheffield DID-ARQ team began by attempting to extract a 'digital fingerprint' from the data, using Polygonal Models and Shape Recognition. Sobel edge detection (from the Image2Learn[18] API) was applied to source images accessed from commonly-shared samples mounted on the project's Medici image library. Line segments were then fitted to edge map data using the expectation-maximization (EM) algorithm (designed to run on multiple cores). Shape recognition algorithms were then applied to polygonal models to identify letters, words, symbols and patterns.[19] We hope to run these algorithms on the NCSA's Petascale High Performance Computing (HPC) machine called Blue Waters or on the NSF Teragrid HPC resources since the computation still takes many hours of computing time; results are beginning to emerge which appear to promise over coming months the potential for identifying a steadily more objective digital fingerprint for some of our hitherto rather shadowy medieval copyists.

3.3 Working with Maps: the Illinois cartography collections

From a preliminary assessment of how British cartographers active between 1685 and 1817 undertook their mapping of the Great Lakes, the University of Illinois DID-ARQ team began to adduce a set of primary initial objectives. The first entailed devising a way of determining the surface area of the irregular shapes found on the maps. Some 40 maps were closely scrutinized, and fresh data generated about the surface area of the lakes. This led to questions being asked as to how much knowledge-sharing appeared to be going on between French and British cartographers during the period under investigation. In turn it became possible to ask whether any unique relationships might be associated between a given national cartographic tradition and (a) specific region(s) of the Great Lakes.

Scholars at Illinois were able to separate out various layers of information going into the maps. These included sundry accounts and sketches made by French traders, armies and explorers, and those made by British travellers and cartographers. It was also possible to identify the diverse ways in which particular 'map houses' tended to integrate their multiple sources into a single map, as the tangible outcome. The automation of map analysis included not only lake area estimation but also the map scale estimation by estimating the graticules with respect to the map latitude and longitude intersections. These analyses allow reporting the map areas in physical units

[18] Image2Learn: http://isda.ncsa.uiuc.edu/Im2Learn/

[19] The computing background to these procedures is covered in Arkin, E.M., Chew, L.P., Huttenlocher, D.P., Kedem, K., Mitchell, J.S.B., "*An Efficiently Computable Metric for Comparing Polygonal Shapes*," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13(3), 1991.

rather than in image pixels which has importance for comparing lake sizes across maps with different scale.

Political and commercial considerations were often significant, although not absolutely determinant: the British assumed control of the regions in question *ca* 1760; the first formal British survey conducted using modern instrumentation took place in 1820. Travellers were able to perform distance calculations using a variety of methods, though these would not be a visible part of any formal survey. Gathering such information was itself often part of larger operations designed to consolidate power or commercial advantage.

The technologies used by these early cartographers gradually improved, especially during the eighteenth century. Qualitative observations became more discriminating, and hitherto inaccurate representations of lake shapes gradually improved (one can quickly establish such exaggerations by comparing the eighteenth-century maps with an overlay from a modern photograph taken by a satellite camera). There were exceptions: connections between the Lakes were sometimes exaggerated. Overall, it was the British maps which tended to evince the most detail.

3.4 Doing the Science: the Illinois cartography collections

When applied to map data covering each of Lakes Ontario, Erie, Michigan, Superior and Huron, pattern recognition algorithms began to generate results permitting some first, tentative interpretations. Horizontal patterns emerged, indicating moments where knowledge was ostensibly being shared between one project and the next. Sometimes there is some kind of sharing going on, though many times there is no manifest 'direct copying' evidenced from one mapmaker to the next. Linear improvement or enhanced coherence of mapping was evinced for some regions (e.g. Lake Ontario), but not for others (e.g. Lake Huron). The question therefore arose: how is one to account for some lakes not being mapped as consistently as others?

One hypothesis had to do with ice. Past explanations for inaccurate-looking maps included ineffective surveying techniques and equipment, varying political control of the region, the desire to represent the Lakes as a passage westwards to Asia, familiarity with some parts of the Lakes but not others due to military and commercial traffic; all of these helped to explain some of the inaccuracies and variations in the maps. In the case of extreme and sustained variation in observed surface area, however, it was postulated that the presence of ice, and the inability of observers to tell the difference between ice and land, might help to explain contemporaneous yet conflicting images of the same lakes across different maps. In particular, icing helped to explain how Lakes Huron and Superior came to vary so greatly in their represented geographical features. This impression was confirmed from a case study focusing on the Northern Straits of Huron. Writing in his journal for 1816, from the south coast of Georgian Bay and in a month usually associated with relative warmth, David Wingfield wrote: '… And on the 2nd *June* I left Lake Huron, still covered with ice for several miles from the shore' (our italics). Even in recent times, a GLERL Digital Ice Chart (for 10 April 1989) registers a heavy preponderance of ice covering much of the area visited by Wingfield. Arial satellite photography confirms the visible confusion that can obtain, despite the precision of modern technologies, between what is land, and what is ice. Mapping what was 'visible' in the eighteenth century could produce some widely divergent results, viz. the very different boundaries for Lake Huron displayed on the maps published by Cary in 1796, Russell in 1801, and Arrowsmith in

1814. Even greater divergences can be registered by comparing Gravelot's map of the same lake from 1746, and Morse's map of 1822.

The implications of these initial findings can be summarized as follows. In terms specifically of Authorship, a great deal of observable variety can be discerned in the maps, both British and French. Some continuities within and between the national traditions can be pinpointed, but they by no means characterize the data. From the standpoint of US Colonial History, it emerges from the data that the French were able to ascertain the shape of lands in Northern Huron sooner than the British or US mapping efforts could. This perhaps points to an exchange of French and Native information resulting in a better sense of the land underneath the ice. Other key factors requiring consideration include Historical Climatology: inconsistencies in surface area across a tradition of 18th-century maps of a given body of water may be indicative of extreme environmental conditions. Data mining applied to 18th-century and early 19th-century map images of other regions, and the examination of trends in surface area recognition may be one way to discover further aspects and regions of interest for the study of 18th-century climate. Moreover, maps of Lake Huron (and the confusion of ice with land) may indicate the influence of relatively longer or harsher winters affecting the Midwest during the Little Ice Age, or in the wake of the 1815 eruption of Mount Tamboro.

3.5 Working with Quilts: the Quilt Index collections

MATRIX hosts the multi-disciplinary Quilt Index, in partnership with Michigan State University Museum and the Alliance for American Quilts (AAQ). This online Index contains images and detailed information on more than 50,000 quilts, dating from the 1700s to the present day. Currently, the quilts are mostly American in origin, though the Index is expanding to include international collections. Quilt images (550 - 1424 pixel-wide JPEG files at 72-150 ppi resolution) have been contributed by multiple museums, libraries and documentation projects for the preservation of American cultural heritage and for research and education in the arts, humanities and social sciences. The set is hosted in MATRIX's open source digital repository, KORA, and available online. Thousands of styles and quilt makers are represented in this dataset, which contains a range of image quality depending on original photography and indicate a wide base of inquiry for humanities questions and challenges for computer science analysis.

Authorship is a complicated notion in the medium of quilts, due to the design and construction techniques inherent in quiltmaking. An individual quiltmaker's authorship may be discerned through the pattern selection, the fabric and color choices, execution of the design in measurement, layout, construction techniques, needlework and craftsmanship, and, most interestingly, in novel deviations from a given pattern. Yet community participation is common across all eras and quiltmaking styles, from traditional to contemporary art quilts, which results in multiple hands across ages, backgrounds, cultures and communities, working on the same quilt. The quilt top (which creates the overall visual layout of the quilt) is likely to be conceived by one person, yet traditions of round robins and block-sharing also exist where, in a group of quiltmakers, each person makes a block and the round robin "owner" pieces them together into a single top. Contemporary art quiltmakers may take old unfinished blocks they have found and work them into a completely new design that calls to and actually contains the previous (and perhaps unknown) maker's block

work. On the other hand, the quilting, or stitches holding the quilt top together with the other two layers (batting and backing), has often been done in traditional quilting bees – community gatherings of women stitching together around a quilting frame. In contrast to the volunteer participation quilting model, niche markets have also evolved of people who specialize in completing various stages of the quilt production process for hire. Roles includes "markers" who specialize in drawing lines in chalk or pencil for the quilting stitching to follow and "quilters" who stitch the three layers together. While this has been true historically in communities, it is especially common practice for contemporary quiltmakers to design and construct the top and then hire or collaborate with a "quilter" to stitch the quilting. Recently a market has developed for quilters who own and use special "long-arm" sewing machines that can accommodate the bulky size of a quilt. Thus, many hands may be involved in the design and construction of a quilt; and it can be difficult to differentiate the contributions of individual authors, even from close visual inspection of the original object by experts.

Quiltmaking relies on a "pattern," which may be of the quiltmaker's invention, but most often is from a design idea or actual template that was either passed down traditionally through families in a community, or obtained from a published source such as a pattern booklet or newspaper column. The quiltmaker's pattern decisions are most evident in the quilt top. Published patterns became much more common starting in the late 1800s, when certain pattern designers mass-produced their patterns and disseminated them through 19th and early 20th century "ladies" magazines, and later in syndicated newspaper columns. Geographically dispersed quiltmakers who were exposed to this media began gaining new patterns and pattern ideas from these sources. Thus in a large test bed of documented historic quilts, the societal rise and influence of mass media should be detectable through the proliferation of quilts that execute patterns disseminated through syndicated columns.

Given these techniques of design and construction, many interesting and relevant questions of authorship arise from visual analysis of quilt images.

- What are the distinct characteristics of an individual quiltmaker or relevant quiltmaking group's choices of pattern selection, fabric and color choices, execution of measurement, layout, needlework and craftsmanship of the pattern design, and novel deviations from traditional patterns?
- To what extent are quilt patterns regional and to what extent national, and do changes in culture and technology meaningfully influence any trends? In a large test bed of documented historic quilts, can the societal rise and influence of mass media be seen through the proliferation of quilts that execute patterns disseminated through syndicated columns? Did the rise of published patterns dilute regional styles?
- Can you trace a regional style as it migrated with a maker to new territories, or detect its absence in certain cultures?
- Can you trace the elements of a published pattern and identify quilts where an individual used that element on an otherwise unique quilt?
- Can the quilts created by quiltmakers from a cloistered family, community, ethnic, or religious group at a particular time period be differentiated from those of other communities, especially those more exposed to mass media? Do patterns of collective authorship reveal cultural practices of the community? If so, can changes in the community's participation in mass culture be found through changes in quiltmaking?
- Can a resurgence or interest in a particular historic cultural community's quiltmaking styles be found in quiltmaking a century later – are there cycles that

are visually manifested in a large database?

3.6 Doing the Science: the Quilt Index collections

For the authorship project MATRIX worked with scholars at the Michigan State University Museum and with researchers from the quilt textile community to identify several challenging questions relating to authorship. The research strategy is two-fold and sequential: first determine relatively salient characteristics of colors, shapes, borders, layouts and patterns within distinct creator groups, then use these results to facilitate automated clustering, and hopefully uncover important new similarities and dissimilarities for a broad range of humanities analysis.

One of the first steps, with the aid of curators, specialists in textiles and folklife material culture, and art historians associated with the Michigan State University Museum, was to analyze the color space and pattern to explore origins and context of a quilt made in Chicora, Michigan in 1926 as a fundraiser for the local Ku Klux Klan. While similar patterns exist, this specific block design and use of colors are not found in published textile documentation. Is this a unique pattern arranged by a single quiltmaker, or is the arrangement of this pattern a symbolic emblem of the Klan organization? Finding similar or identical quilts through this image data analysis project will help reduce the search for relevance and potentially inform demographic estimates of latent racial bias over time in important new historical information and meaning. The collaborators are exploring two foundational research questions that will impact how quilts and quiltmaking can contribute to addressing these issues: 1) How can fabric colors be used to identify or to narrow the search for novel objects and shared or similar authorship among a large, diverse dataset of quilt images? 2) How can shape pattern identification be used to find novel or similar patterned objects?

Humanities experts identified the choice of blue colored fabric and redwork embroidery in the white blocks as fundamental to the meaning of this quilt; consequently computer scientists isolated the “indigo blue” used in this quilt. MSU computer science researchers developed algorithms to conduct pixel-by-pixel analyses of the quilt images, converting the standard RGB (red, green, blue) levels to HSL (hue, saturation, lightness) values for greater evaluation accuracy. HSL was chosen because various shades of indigo dye have the same hue and saturation, but varying degrees of lightness. In addition, fabric fading over time primarily appears to affect the lightness and saturation. Choosing this color space thus allowed adjustment and weighting in the values of hue, saturation and lightness to figure the nearness to an ideal indigo color. The algorithm was tested and revised on a controlled set of 50 known images and ultimately achieved 89% accuracy within that defined sample set. This method and algorithm was most successful with quilts that had more indigo color present, but began to lose accuracy as less indigo was present. When this algorithm is applied across the entire Quilt Index, we expect this high accuracy rate will diminish as the variation in image quality, color calibration, and consistency increase. The simple ability to search images by color will be a highly valued contribution to the Quilt Index since only broad categories of color are routinely recorded in the metadata, but not finer-grained gradations and certainly not to the level of specificity that computer analysis can provide. Not only can scholars trace trends in color use, but many of the users of the Quilt Index are quiltmakers and designers who would greatly value the ability to find quilts based on colors used.

While color was a successful identifier in some research questions, more complex pattern recognition was required in others. Crazy quilts date back to the Victorian period and, as the name implies, they look somewhat "crazy", constructed in an unpredictable arrangement of colors and odd shapes.[20] The pieces used to construct the top are irregular in shape and size; incongruent pieces of fabric are arranged randomly in the quilt top – the front design surface most commonly viewed. This random arrangement is very easy for a human to recognize, but much harder to describe in accurate terms that can be translated to an algorithm; this challenge makes for an interesting computer science problem. Development of a crazy quilt algorithm was accomplished by a student working with the NCSA researchers at UIUC. The resulting algorithm divides the image space of the quilt into distinct regions and compared the arrangement of colored pieces in each region to the arrangement of color pieces in each other region to determine the extent of similarity across the quilt. The less similarity among regions, the more likely the quilt might be a "crazy" quilt. MSU utilized the High Performance Computing Center to run the crazy quilt algorithm on the entire corpus of more than 56,000 images. The resulting categorization of quilts as "crazy" or "not-crazy" points to a number of new directions. First, while researchers are currently working on improving the results, and what has proven interesting is not simply the correct results but those quilts deemed to be "inaccurate" results. Many quilts that were incorrectly categorized as "crazy" appear visually related to crazy quilts. These could prove, with further research by quilt scholars, to have a relationship to the development and subsequent transformations of the crazy quilt style or to the influence of this chaotic visual manifestation on subsequent pattern development and quilt art production.

This complex pattern recognition will be used to analyze the patterns of Amish quilts, to try to identify previously unassigned Amish made quilts and to uncover trends in the influence of Amish-made quilt designs on mainstream American culture. The trained human eye can often easily identify traditionally patterned Amish quilts, which, depending on region, tend towards either verticality or centered medallions, overall symmetry, and solid darker colors when compared to non-Amish made quilts during the same time period.[21] Yet the designs can vary greatly, making seeking such quilts with visual parameters a challenge for computational analysis. Currently quilt scholars are working with MSU researchers to define the salient features of Amish quilts from several mid-western regions in preparation for developing an algorithm that will build on the color extraction algorithm and perhaps the symmetry/region similarity algorithms developed for "indigo" and "crazy" quilts.

4 Extrapolations: across the consortium and its datasets

As the project moved forwards, algorithms (polygonal matching and edge detection, 'ball-based segmentation', classification) were applied successively to the Maps collection, and then to the Quilts collection. Once again, early results appeared to confirm the viability and pertinence of these procedures. For example, ball-based segmentation analysis applied to the Maps collection revealed characteristic

[20] Jan Przybysz, "The Victorian Crazy Quilt as Comfort and Discomfort" in *The Quilt Journal*, Volume 3:2, 1994. (Kentucky Quilt Project, Inc: Louisville, KY)

[21] Rachel Pellman, "Distinctive Features of Amish Quilts" in Pellman and Pellman *A Treasury of Amish Quilts* (Good Books: 1990), pp. 32-33.

differences between 17th- and 18th-century maps produced respectively by French and English/American cartographers. Useful conclusions were also adduced from applications of the algorithms to sub-sections of the Quilts database (in particular, to so-called 'crazy' quilts evincing no readily discernible pattern and to the color space of indigo-blue).

4.1 Looking towards the future

Just fifteen months or so of collective endeavour have already produced some tentative theoretical conclusions as well as some firm results and outcomes. The project has demonstrated the viability of this kind of collaborative endeavour, as well as the added value secured through the sharing mutual trials of algorithms and procedures across disciplinary boundaries. Some cogent ideas are emerging, too, for future developments that go well beyond the initial DID-ARQ project brief.

4.1.1 Froissart MSS corpus

Some initial work on illuminated (ornamental) letters and paragraph markers in the manuscripts has begun, but at time of writing has had to be deferred pending completion of a DHQ paper[22]. Provisional results confirm that there is ample scope for detailed letter analysis to be undertaken by the project partnership, within which algorithms for scribal hand classification are already undergoing further development and refinement. It is anticipated that more sensitive algorithms for the analysis of ductus may yet be devised, leading to the elaboration under a fresh project brief of more dynamic, 'kinetic' approaches to machine analysis of scribal and book hands which in turn will lead to greater insights into the structure and functionality of the Parisian book trade.

4.1.2 The maps corpus

Based on the preliminary analyses of the maps of the Great Lakes printed between 1650 and 1850, we believe that there is a good deal of suggestive information on climatic conditions. The team has begun discussions with the School of Earth, Society, and Environment, and the Department of Atmospheric Sciences at UIUC to joint forces in understanding global warming patterns. We anticipate that bringing together the results from historical maps and the observations from more recent climatological observations of the Great Lakes will reveal any periodic climatic patterns over time and enable us to understand the relationships to cartography and trade.

4.1.3 The quilts corpus

The initial tests for pattern matching of crazy quilts and the color indigo-blue promises to open up new ways of building user interfaces (based on pattern matching and color matching) as well as new ways for quilt scholars to do their work more quickly and efficiently. Further research questions include refining the color and

[22] Natalie Hansen, T. Shaw, H. Tennison, A. Hedeman, and P. Bajcsy, "Cyber Connoisseurship: The Application of Cyber Tools to Aid Understanding of the Medieval Parisian Book Trade", under consideration for *DHQ*

pattern algorithms and applying them in concert to search for additional instances of the block pattern arrangement found in the Chicora KKK quilt. In addition, we are now working toward pattern matching of Amish quilts to determine the spread, migration, and transformations of typically Amish patterns. Quilt scholars are presently collaborating with us on determining a feature set for Amish quilts.

4.2 General Conclusions

With so much virtual content available to examine and with such subtle nuances to isolate and compare, the ability to cluster materials that appear to contain similar characteristics is proving to be of great benefit. In addition to the use of computer algorithms such as Sobel edge, segmentation, classification and Polygonal Shape matching, and through the deployment of more straightforward tools such as email and ooVoo-enabled videoconferences, the project's collaborative sharing of tools has enabled team members to share, view, manipulate, discuss and work on the three varieties of image (maps, quilts and manuscripts) in real-time and via live discussions and debate. Removal of geographical boundaries and distance between experts, which can slow down the research process, has helped to establish a platform for the efficient sharing and development of emerging ideas and concepts. The project has also provided a shared platform and toolset for making observations about images which can be read and commented upon by others, and used by future projects. Not the least valuable dimension to the project is the opportunity it has begun to provide for graduate training.

Despite the usefulness of the e-Science tools deployed for shrinking the barriers between sites and team members, the project has funded to great advantage visits to the USA by both members of the Sheffield team, to refine and develop on-site at NCSA some of the more complex technologies and algorithms developed during the project's recent months, and to participate in the June 2011 Washington DC conference. A workshop will also be held during the same month at Sheffield's Humanities Research Institute, to roll out the project's findings and outcomes, and to offer food for thought to potential applicants to DID2.

We would like to provide some final thoughts about some fundamental but perhaps little known advantages of working with digital artifacts. Traditionally, scholars have travelled to the manuscripts, maps or quilts, visiting the collections and research libraries in which these extremely valuable artifacts are housed, and consulting them on site. So precious are some of them, and so sensitive to damage from daylight, that conservators and librarians have been increasingly reluctant to provide scholars with more than the briefest spell of direct contact with the artifacts themselves, the remainder of the inspection being of necessity confined to a black-and-white microfilm or photographs. Time and availability of staff is a related issue. Some libraries never allow scholars to view the originals at all.

Whilst it remains possible to take careful notes, to order digital photographs of particular items and to move on to the next library or libraries to make fruitful comparisons between the various artifacts, the elements of at least one part of our corpus find themselves scattered today across Western Europe and North America, with Froissart manuscripts to be found in Chicago, Austin (Texas), but also Brussels, Glasgow, London, Besançon, Paris, Toulouse and Brussels (one turned up on 8 April 2011 in a saleroom in Paris with a reserve price of 300,000 €). The advent of high-resolution digital photography and grid computing has provided the conditions for a

revolution in the study of corpora like ours: the manuscripts, maps or quilts can now 'come' to the scholar or team of scholars, virtually and over the web. In the manuscript field, the internet has opened the door to large-scale collections such as those held by the British Library (*Online Catalogue of Manuscripts*), the Bibliothèque nationale de France (*Gallica*) and the Albertine in Brussels (*Belgica*), though usually at resolutions of only 300 or 150 dpi, and without the collaborative, interactive or federative tools that scholars would like to have access to. Rarely, moreover, does one find digital copies of *entire* manuscripts made available over the web, still less of an entire corpus (materially scattered, but assembled in virtual reality). Rarer still are intensive research partnerships built around such corpora, facilitated and encouraged by enlightened librarians and assembled in virtual form from around the world to form the object of comparative and collaborative scholarship conducted by international teams of academics bringing to the table complementary fields of academic expertise. That is what makes the *DID-ARQ* project so unusual, so innovative and so exciting to be a part of.

In addition, the use of computer algorithms allows extracting information that would not be possible to extract by visual inspection. For example, converting images of manuscripts into multiple color spaces leads to new representations and better understanding of the color composition. Similarly, estimating lake area by counting pixels would be almost impossible considering the amount of labour required. In the same vein, labelling thousands of quilt photographs as crazy and non-crazy, or searching for the color indigo by hand would not be well-spent time and resource.

Finally, the traditional humanities research has been known to report conclusions derived from a small volume of artifacts. The automation of image mining studies open a new avenue for significantly increasing the confidence (or statistical significance) of these reported conclusions as the results become available over a large volume of artifacts.

5 Acknowledgment

We would like to acknowledge the NSF/NEH/JISC Digging into Data (NSF grant ID: 1039385), and NSF ITS 09-10562 EAGER grants for the support of this work.

The work presented in this paper is a collective effort of the following contributors from the University of Sheffield, NCSA, UIUC, MATRIX, MSUM, and University of South Carolina: Peter Ainsworth, Michael Meredith[23]; Peter Bajcsy, Rob Kooper, Luigi Marini, Tenzing Shaw[24]; Anne D. Hedeman, Robert Markley, Michael Simeone, Natalie Hanson, Heather Tennison, Simon Appleford[25]; Jennifer Guiliano[26]; Dean Rehberger, Justine Richardson, Matthew Geimer, Zachary Pepin, Steve M. Cohen[27]; Marsha MacDowell, Mary Worrall, Amanda Sikarskie, Beth

[23] The University of Sheffield, UK
[24] National Center for Supercomputing Applications, 1205 W. Clark Street, Urbana, IL 61801
[25] University of Illinois at Urbana-Champaign, Urbana, IL 61801
[26] Center for Digital Humanities, University of South Carolina, Columbia, SC 29201
[27] MATRIX: The Center for Humane Arts, Letters, and Social Sciences Online, Michigan State University, East Lansing Michigan, 48824

Donaldson[28]; Alhaad Gokhale[29]. We would like acknowledge the DID-ARQ team by citing Aristotle, *Metaphysica,* "The whole is more than the sum of its parts."

The DID-ARQ project description and its visual materials pertinent to each data set can be found on the three projects web sites:

NCSA/UIUC: http://isda.ncsa.uiuc.edu/DID/index.html

MSU: http://projects.matrix.msu.edu/did/

University of Sheffield: http://www.hridigital.co.uk/did

[28] Michigan State University Museum, Circle Drive, East Lansing, MI 48824

[29] Indian Institute of Technology, Kharagpur, India