Das CAMENA-Projekt: Erfahrungen mit der Digitalisierung alter Drucke in lateinischer Sprache

Hinter der Bezeichnung CAMENA stecken eigentlich vier verschiedene DFG-Projekte, die aber thematisch verwandt, eng verzahnt und als einheitliches Angebot konzipiert sind. Hauptinhalt der Projekte ist die digitale Veröffentlichung lateinischer Literatur aus dem deutschen Sprachraum von ca. 1450-1750, der bei weitem meisten Digitalisate sind dichterische Werke, der Anteil der Fachprosa ist aber im Steigen.
Diese Materie mag zunächst als randstädige Kuriosität erscheinen. Betrachtet man aber die Politik der führenden Förderinstitutionen, kommt man um die Erkenntnis nicht umhin, daß Projekte in diesem Bereich höchst erwünscht und einschlägige Anträge erfolgversprechender sind als in anderen Themenbereichen. Obwohl auch die Digitalisierungstöpfe der DFG nicht mehr so voll sind wie zu Beginn der Programme, hat die DFG seit 1999 (rechnet man die laufenden Phasen ein) etwa achthunderttausend Euro für die CAMENA-Projekte bewilligt.
Diese Großzügigkeit hat mehrere Gründe:
- Angesichts des immer stärker beschränkten Zugangs zu alten Drucken ist der Sinn der Digitalisierung bei diesen Materialien unmittelbar einsichtig, besonders da gerade die zahllosen neulateinischen Werke nur in wenigen Einzelfällen in Neuausgaben vorliegen, obwohl in den Drucken bis 1700 das Lateinische eindeutig dominiert.
- Die Digitalisierung alter Drucke bedeutet zugleich die Schonung der Originale. So werden die Originale der im CAMENA-Projekt digitalisierten Werke nur noch ausgegeben, wenn der Nutzer plausibel macht, daß sein besonderes Interesse mit dem Digitalisat nicht befriedigt werden kann (ein Fall, der meines Wissens noch nicht eingetreten ist).
- Das internationale Forschungsinteresse an den neulateinischen Quellen ist seit längerem im Steigen, die Forschungslage wegen des schwierigen Zugangs aber noch weithin schlecht.

Eine Bibliothek, die über umfangreiche Altbestände verfügt, hat also hervorragende Erfolgsaussichten, wenn sie diese mit einem intelligenten Konzept und begleitender fachwissenschaftlicher Kompetenz digitalisieren möchte.
Ich möchte deshalb im folgenden das CAMENA-Modell und unsere Erfahrungen im Umgang mit der Materie vorstellen, in der Hoffnung, daß dadurch vergleichbare Projekte erleichtert und ermutigt werden.

Der Editionstyp, den das CAMENA-Projekt verfolgt, ist meines Wissens noch singulär: CAMENA bietet nebeneinander Bildscans der Originale und Volltextabschriften im XML-Format, die mit den Originalseiten verlinkt sind. Da Bibliotheken oft reine Bildeditionen bevorzugen, deren Herstellung und Publikation weniger aufwendig ist, möchte ich zunächst die Entscheidung für die Volltextedition begründen.

Wie u. a. Klaus Graf zu Recht – auch in der InetBib-Liste – mehrfach kritisiert hat, sind reine Bildeditionen, wie sie leider von mehreren Bibliotheken reichlich mechanisch und mit geringem Erschließungsaufwand erstellt werden, für die Nutzung nahezu wertlos; selbst bei guter Erschließung bleiben sie an Komfort deutlich hinter dem Original zurück. Der Nutzer leidet entweder unter der schlechten Bildqualität oder unter langen Ladezeiten. Bei Materialien von vorwiegend dekorativem Wert ist natürlich ein Volltext weniger sinnvoll, aber im Falle der neulateinischen Literatur ist das kunsthistorische Interesse marginal. Eine reine Bildedition bliebe hier an Gebrauchswert hinter dem Original zurück und würde sogar das wichtigste Anliegen der Nutzer - nämlich die corpusübergreifende Wort- und Zeichensuche - nicht bedienen. Warum also überhaupt die Bildpublikation? Wenn Einigkeit in der Fachwissenschaft bestünde, wie neulateinische Texte heute zu edieren sind, und wenn es möglich wäre, flächendeckend fehlerfreie oder wenigstens fehlerarme Volltexte der lateinischen Originale anzubieten, wären die Bilder für die meisten unserer Nutzer in der Tat verzichtbar. Beides ist leider nicht der Fall.

Die Scans des CAMENA-Projekts werden im Hause von einem eigens beschäftigten technischen Mitarbeiter erstellt. Diese Lösung ist nach unserer Erfahrung einfacher, flexibler und zumindest nicht teurer als die zunächst von der DFG propagierte Vergabe der Scanarbeiten an externe Dientleister. Derselbe technischen Mitarbeiter veranlaßt nach der Veröffentlichung der Bildscans die Registrierung der Digitalisate als eigene Titel im Mannheimer OPAC (und darüber in den Verbundkatalogen SWB und KVK) und in der neulateinischen Online-Bibliographie von Herrn Prof. Sutton (Philological Museum) sowie die Versendung der Vorlagen für die Volltexterfassung.

Da auch gute OCR-Software den Variationen und Unregelmäßigkeiten der handgegossenen und handgesetzten Typen der Bücher vor 1700 nach unserer Erfahrung nicht gewachsen ist, müssen die Volltexte von Hand abgeschrieben werden. Versuche, diese Arbeit von lateinkundigen Studenten erledigen zu lassen, mußten wir rasch wieder abbrechen. Die Hilfskräfte waren leider nicht in der Lage, die Texte hinreichend schnell oder auch nur hinreichend korrekt zu erfassen. Es hat sich - bei ebenso guter Qualität- als wesentlich kostengünstiger erwiesen, die Texte von lateinunkundigen (chinesischen) Typisten erfassen

zu lassen. Die Typisten markieren nicht nur schwer lesbare Stellen, sondern führen auch nach Anweisungen der Projektleitung einfache layoutbasierte XML-Kodierungen durch. Der abgeschriebene Text wird dann mit entsprechender Software orthographisch normiert - was im Interesse der Absuchbarkeit unabdingbar ist - und morphologisch analysiert. Die nicht analysierbaren Formen (derzeit je nach Vorlage etwa 3-7 Prozent) werden in Form einer Fehlerliste ausgeworfen und manuell korrigiert; der Wortschatz der Erkennungssoftware wird dabei im Zuge der Korrekturarbeiten stetig erweitert. Das Resultat ist zwar kein fehlerfreier, aber immerhin ein gut lesbarer, mit guten OCR-Scans moderner Werke vergleichbarer Volltext, aus dem Inhaltsverzeichnissse und Indizes je nach angewandter Kodierung extrahiert werden können. Die noch vorhandenen Fehler kann der Nutzer durch die Verlinkung mit den Seitenbildern leicht identifizieren. Da die XML-Texte auch ins HTML-Format konvertiert vorgehalten werden, ist jede Zeichenkette des Textcorpus von Internet-Suchmaschinen absuchbar. Mit der lokalen Suchfunktion sind darüberhinaus auch Suchen nach bestimmten Lemmata oder Kollokationen innerhalb des ganzen Corpus oder in ausgewählten Teilstücken vorgesehen; zudem kann aus den Volltexten eine OAI-Registrierung extrahiert werden (die beiden letztgenannten Funktionen sind leider noch nicht implementiert). Angesichts der vollständigen Absuchbarkeit und der automatischen Erstellung von Inhaltsverzeichnissen erübrigt sich eine Verschlagwortung. Unsere Nutzerstatistik zeigt deutlich, daß die Veröffentlichung von Volltexten in einer für Suchmaschinen lesbaren Form die Nutzbarkeit entscheidend fördert: Die meisten Internetnutzer navigieren immer noch hauptsächlich über Google-Suchen. Trotz der Sprachbarriere ist nach der Edition umfangreicher Volltexte die Zahl der Zugriffe exponentiell angestiegen, zuletzt auf die für ein lateinischsprachiges Angebot sehr beachtliche Zahl von 700-1000 visits pro Tag. Der gelegentliche Druckfehler beeinträchtigt besonders bei einem großen Corpus den Erfolg von Suchvorgängen kaum noch. Freilich muß in Anbetracht der großen Mengen "schmutziger" Daten von einer einheitlichen Datenaufbereitung Abstand genommen werden. Aufgabe der Projektleitung muß es sein, über den jeweiligen Grad der Aufbereitung nach inhaltlicher Gewichtung zu entscheiden und im Kontakt mit dem Fachpublikum für eine Beteiligung der Nutzer an der editorischen Arbeit zu sorgen.

Für die Qualität der Abschriften ist natürlich neben der Softwarepflege vor allem die Arbeit der Datenerfasser entscheidend. Unsere Erfahrung ist, daß es sich auszahlt, sich eine Firma "heranzuzüchten"; denn trotz fehlender Sprachkenntnis steigt die Erfassungsgenauigkeit mit dem erfaßten Volumen stark an. Außerdem ist es für die Erfasser attraktiv, eine gewonnene Sonderfertigkeit wie die des Erfassens alter Drucke in fremden Sprachen zu pflegen; sie sind

daher im Zweifelsfall bereit, günstigere Preise anzubieten. Natürlich wird von Seiten der DFG nur der billigste angebotene Tarif bewilligt. Ein zuverlässiger Datenerfasser muß also willens und in der Lage sein, die Konkurrenz zu unterbieten. Er sollte außerdem bereit sein, schwierige Typen anzulernen (neben Fraktur z. B. die oft stark abbreviierten griechischen Texte). Vor allem aber muß er in der Lage sein, zuverlässig Kodierungen nach Anweisungen der Projektleitung einzufügen. Diese Kodierungen sind für die spätere Verwendbarkeit der Volltexte von großer Bedeutung und von Projektseite nur unter erheblichen Mehrkosten nachträglich einzufügen. Es ist die bei weitem rentabelste Lösung, wenn sich ein fachkundiger Projektmitarbeiter vor der Erfassung intensiv mit den verwendeten Layoutsignalen auseinandersetzt und die inhaltlich entscheidenden Merkmale zur Kodierung ausweist.

Fachwissenschaftliche Kenntnis ist somit, bei aller computertechnischen Perfektionierung, unumgänglich. Das beginnt bei der intelligenten Auswahl der Vorlagen. Nutzer sind nun einmal nicht bestands-, sondern inhaltsorientiert. Eine flächendeckende Dokumentation eines Einzelbestandes geht an ihren Interessen vorbei. Vielmehr muß geprüft werden, ob die Vorlagen inhaltlich relevant, philologisch maßgeblich und nicht bereits anderweitig neu ediert sind. Wenn man sich für eine reine Bildedition entscheidet, muß diese von Fachleuten sorgfältig erschlossen werden; bei der meist zu bevorzugenden Volltextedition muß ein Fachmann entscheiden, welche Textmerkmale für die informatische Auswertung entscheidend sind. Die Kontakte zu den Nutzern, die ja nach Möglichkeit in die Korrekturarbeiten einzubinden sind, kann ebenfalls ein Fachmann am besten knüpfen und pflegen; und schließlich muß die Entscheidung, welche Teile eines größeren Corpus besonders zu pflegen sind, inhaltlich zu verantworten sein.

Die Richtlinien der Deutschen Forschungsgemeinschaft für ihre Digitalisierungsprojekte waren in der nun zu Ende gehenden Experimentierphase absichtlich vage. Die Projekte waren damit in der Wahl der technischen Lösungen vorwiegend auf sich selbst gestellt, was zweifellos viele potentielle Antragsteller von komplexeren Herausforderungen, für die keine erprobten Lösungen vorlagen, abgeschreckt hat. Wenigstens für den zukunftsträchtigen Bereich der Retrodigitalisierung alter Drucke in lateinischer Sprache können wir das Erfassungsmodell der CAMENA-Projekte mittlerweile als erprobt und effizient empfehlen. Jede Bibliothek, die über entsprechende Altbestände und wenigstens eine fachkundige Kraft verfügt, hat damit gute Aussichten, in diesem Bereich erfolgreiche Drittmittelanträge zu stellen. Sie tut damit nicht nur ihren Finanzen und ihren Nutzern, sondern vor allem ihren Altbeständen den größten Gefallen.