

Die digitalen Monumenta Germaniae Historica

I. Projektrahmen

Die Bemühungen der Monumenta Germaniae Historica (MGH) auf dem Gebiet der elektronischen Publikation und der Digitalisierung stehen seit Mitte 2004 ganz im Zeichen der digitalen MGH (dMGH).¹ In diesem von der DFG geförderten Projekt, das die MGH in Kooperation mit dem Münchener Digitalisierungszentrum der Bayerischen Staatsbibliothek (BSB) durchführen, werden bis voraussichtlich 2010 sämtliche MGH-Editionen digitalisiert und frei im Internet zugänglich gemacht.

Bei den dMGH handelt es sich keineswegs um das einzige elektronische Vorhaben der MGH, doch können im Rahmen dieses Berichts andere erwähnenswerte Projekte nicht näher vorgestellt werden. Immerhin sei kurz festgehalten, dass binnen Jahresfrist die vollständige Digitalisierung der Zeitschrift „Neues Archiv“ sowie deren Nachfolgezeitschrift „Deutsches Archiv“ mit frei verfügbaren Volltexten abgeschlossen sein soll. Die Monumenta sind also derzeit dabei, in großen Schritten den überwiegenden Anteil ihrer gedruckten Publikationen online zugänglich zu machen. Der Hauptanteil davon entfällt naturgemäß auf die Editionsbinden, die durch die hier zu behandelnden dMGH abgedeckt werden.

Die MGH haben in den über 180 Jahren ihres Bestehens in mehreren hundert Bänden die größte Sammlung schriftgebundener Quellen zur mittelalterlichen Geschichte in kritischen, methodisch prägenden Ausgaben herausgebracht und als Vorbild für ähnliche Unternehmungen in vielen anderen europäischen Nationen gedient. Da das mittelalterliche Kaiserreich auch Italien, Burgund, Böhmen, die Schweiz sowie die heutigen Beneluxländer umfasste und im Osten bis weit in den heute polnischen Raum ausgriff, ist diese für die deutsche Geschichte wichtigste Sammlung auch im europäischen Maßstab und weit über den Kreis der Fachhistoriker hinaus für Philologen, Juristen, Theologen und Kulturwissenschaftler von zentraler Bedeutung. Dies gilt auch und insbesondere für die ‚vornationalen‘ Epochen wie die Merowinger- und Karolingerzeit.

In den Bibliotheken sind die MGH meist als Präsenzbestand aufgestellt, also für den einzelnen Forscher nicht frei verfügbar. Der Gesamtbestand war bislang nur mangelhaft durch einen Registerband von 1890 erschlossen. Durch die Bereitstellung der dMGH wird allerdings nicht nur die Zugänglichkeit und die Erschließung verbessert. Die Volltextfassung erbringt auch einen Mehrwert gegenüber der gedruckten Ausgabe. Die freie Präsentation der MGH-Bände im Internet verleiht der Ursprungsidee der MGH, die Geschichtsschreiber des Mittelalters als ein nationales Erbe mit modernen Mitteln zu erschließen und zugänglich zu machen, eine neue Dimension. Bereits die Erfahrungen des ersten Projektstadiums, in dem lediglich die Bilder der Druckseiten, nicht aber die Volltexte zugänglich waren, zeigen, dass sowohl die Forschung wie auch der akademische Unterricht im In- und Ausland in erheblichem Maße auf das Gesamtkorpus zugreifen.

Die digitale Konversion, Erschließung und die Bereitstellung wird in Kooperation mit dem Digitalisierungszentrum der Bayerischen Staatsbibliothek durchgeführt und umfasst den gesamten Inhalt der Editionsbinden, also auch Einleitungen, kritische Apparate und Register. Die Editionen werden in Form von digitalen Faksimiles der Buchseiten mit dem recherchierbaren, im Hintergrund vorliegenden Volltext zur Verfügung gestellt. Die zukünftig erscheinenden Editionsbinden werden, unter Einhaltung einer kurzen Sperrfrist, zwischen Erscheinen der gedruckten und der digitalen Edition („moving wall“) in das Angebot integriert werden.

II. Projektarbeit

Wie sah bzw. sieht konkret die Projektarbeit aus? In einem ersten Arbeitsschritt wurden die Bände komplett eingescannt. Der Gesamtumfang beträgt insgesamt mehr als 350 Bände mit mehr als 160 000 Seiten. Da diese Bilder nicht nur zum direkten Lesen am Computermonitor gedacht sind, sondern vor allem auch als Grundlage für die Volltextfassung mittels automatischer Texterkennung (OCR) dienen sollen, waren hier besondere Qualitätsanforderungen an die Vorlagen und die Scans zu stellen.

Es ist freilich gar nicht so einfach, von den alten Bänden brauchbare Exemplare aufzutreiben, die für eine Digitalisierung mit anschließender OCR geeignet sind. Oftmals sind die Bände durch Wasserschäden, Stockflecken oder Vergilbung so angegriffen, dass die Scans hiervon kaum zu gebrauchen sind. Auch auf die zahlreichen Nachdrucke konnte aus zweierlei Gründen nicht zurückgegriffen werden: zum einen war die Qualität der fotomechanischen Nachdrucke in der Regel nicht ausreichend, sodass eine auch nur halbwegs fehlerfreie Texterkennung nicht mehr gewährleistet war. Zum anderen aber entsprechen die Nachdrucke nicht immer in jedem Detail den Originalausgaben, da neben den ohnehin aktualisierten Titelblättern oft auch die Abbildungen an anderen Stellen eingebunden und in Einzelfällen auch kleinere Korrekturen vorgenommen wurden (beispielsweise Verbesserungen falsch gedruckter Seitenzahlungen).

Fand sich dann ein gut erhaltenes Exemplar, so konnte es dennoch sein, dass es durch Eintragungen von Korrekturen und Ergänzungen durch wohlmeinende Benutzer ebenfalls für unsere Zwecke nicht mehr brauchbar war. Einige Ausgaben konnten tatsächlich nicht in den Beständen der MGH oder der Staatsbibliothek in hinreichender Qualität gefunden werden. Hier konnten teilweise andere Münchner Bibliotheken hilfreich zur Seite stehen, in einem Fall wurde gar auf die Universitätsbibliothek in

The screenshot shows the MGH website interface. At the top, there is a navigation bar with links for 'Home', 'Nutzungsbedingungen', 'Impressum', and 'Hilfe'. Below this is a search bar with buttons for 'Reihenübersicht', 'Erweiterte Suche', 'Treffer', 'Drucken', 'Blättern', 'Lesezeichen', 'Vollansicht', and 'Html'. The main content area displays the title 'Dagobert I. ernennt seinen Thesaurar Desiderius zum Bischof von Cahors. (630) April 8 (Ostern), -' and the text of the document. The text is presented in a digital edition format with line numbers (10, 15, 20, 25) on the left. The text includes references to various editions and manuscripts, such as 'Vita s. Desiderii: Paris, BN, Ms. lat. 17002, fol. 209v. Kopie 9/10. Jh. (V1) - Kopenhagen, Kongelige Bibliotek, Ms. Thott 136 fol., fol. 7r-8r. Kopie 14. Jh. (V2) - Paris, BN, Ms. lat. 11762, fol. 237r-237bisr. Kopie 17. Jh. (V3)'. The interface also includes a table of contents on the left and a search bar at the top.

Abbildung 1: Anfang einer typischen Urkundenedition: der hinterlegte Textteil demonstriert das optische Hervorheben eines Suchbegriffs

Tübingen zurückgegriffen. In Ausnahmefällen wurden auch Bände antiquarisch nachgekauft. Dennoch blieb insbesondere bei den älteren Bänden ein mitunter erhebliches Maß an Nachkorrekturen zu leisten. Das Einscannen der Bände erfolgte überwiegend im Haus bei der BSB mit einer Qualität von 600 dpi, die für die OCR geeignet ist. Im Rahmen des personell, finanziell und zeitlich Möglichen wurden sodann erforderliche Bildbearbeitungen vorgenommen. Insbesondere wurden die Bilder in einem halbautomatischen Verfahren geradegerückt und störende Flecken entfernt.

Bei der Auswahl der Bände und der Vorbereitung des Scanvorgangs wurde sogleich eine rudimentäre sachliche Erschließung der Texte vorgenommen: Das Inhaltsverzeichnis wurde erfasst, bibliographische Angaben wurden notiert und für jede zu scannende Seite wurde ein entsprechender eindeutiger Dateiname vergeben und die Zuordnung der gedruckten Seitenzahl zum Dateinamen des Scans vorgenommen. Diese Informationen werden in einer Datenbank vorgehalten, die auch die Grundlage der weiteren Verarbeitungsschritte bis hin zur Präsentation im Internet bildet. Für die Präsentation der reinen Bilddateien konnte auf das allgemeine Bereitstellungssystem der BSB zurückgegriffen werden. Bereits dieses Angebot, das im wesentlichen Mitte 2005 fertiggestellt war, erfreut sich großer Beliebtheit, was sich in den beträchtlichen Zugriffszahlen auf die Webseite niederschlägt.

Die Image-Digitalisierung stellt aber nur den ersten Schritt dar. Der zweite und weitaus aufwändigere Arbeitsschritt besteht in der Texterfassung und der Entwicklung der Präsentationssoftware mit Volltextsuchfunktion. Die Volltexterfassung, mit der ein externer Dienstleister beauftragt wurde, begann im Frühjahr 2005. Sie wird bis voraussichtlich Mitte 2010 laufen. Plangemäß wurden im ersten Projektabschnitt bis Mitte 2006 die Bände der Diplomata- und Epistolae-Reihen erfasst. Derzeit läuft die Erfassung des mengenmäßig größten Teils, der Scriptorum-Bände. Im abschließenden Projektabschnitt ab 2008 werden dann die restlichen Bände (Leges, Antiquitates und die weiteren Reihen) folgen.

Es gehört zu den besonderen Qualitätsmerkmalen der MGH-Editionen, dass den Herausgebern die Freiheit gelassen wird, Editionstext, kritischen Apparat, Sachanmerkungen und Register nach den jeweiligen Anforderungen des zu edierenden Textes zu gestalten. Die heterogenen Erscheinungsformen der Einzelbände gestatten also bei der Volltexterfassung kein pauschalisiertes Vorgehen, wenn auch im Rahmen dieses Projektes nicht generell alle Besonderheiten des Druckbildes der jeweiligen Bände adäquat in eine Textform umgesetzt werden können. Hier musste ein Mittelweg gefunden werden. Ziel ist es, so viele semantische Auszeichnungen zu übernehmen, wie anhand optischer Kriterien mit programmgesteuerten Methoden aus dem Druckbild zu eruieren sind.

Ein Beispiel anhand eines typischen Diplomata-Bandes möge dieses Vorgehen erläutern (vgl. Abb. 1): Der Beginn einer Urkundenedition ist an der zentrierten Urkundennummer in Fettdruck zu erkennen. Etwaige weitere Vermerke („unecht“ etc.) sind in der selben Zeile rechtsbündig zu finden. Darauf folgt als kurzer Absatz das Kopfrege. In der unmittelbar folgenden Zeile findet sich rechtsbündig die Datums- und Ortsangabe. Es folgt ein Zwischenraum. Die folgenden Absätze des Kopfteils sind insgesamt links eingerückt. Auf den Kopfteil folgt der eigentliche Urkundentext. Vergleichbare Angaben lassen sich für die Apparate und die anderen Bestandteile der Edition machen. Nicht berücksichtigt werden können hier allerdings rein inhaltliche Kriterien. So ist es nicht möglich, im Rahmen des dMGH-Projektes etwa die einzelnen Formularbestandteile der Urkunden auszuzeichnen, da dies nicht allein mit maschinellen Mitteln möglich ist, sondern den manuellen Eingriff eines verstehenden, qualifizierten Bearbeiters erfordert. Zwar können nach dem oben angedeuteten Verfahren pro Reihe

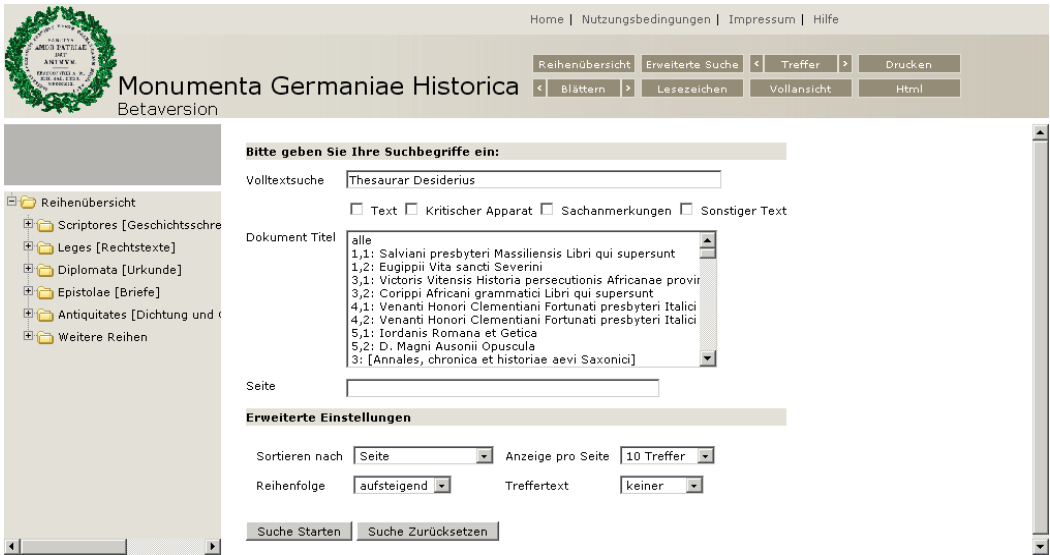


Abbildung 2: Suchformular der dMGH

generelle Erfassungsvorgaben erarbeitet werden, aber zusätzlich ist jeder Einzelband zu analysieren und in Form eines Pflichtenheftes für den externen Dienstleister zu beschreiben.

Ein Nebeneffekt der Erfassung mit OCR ist, dass zu jedem Wort auch seine genaue Position auf dem Image bekannt ist. Diese Informationen werden in der Regel nach der OCR verworfen, da sie nicht weiter benötigt werden. Für das dMGH-Projekt stellen sie aber einen zentralen Bestandteil der zu erfassenden Daten dar. Die Koordinaten, die zu jedem Wort das umgebende Rechteck definieren, lassen auch in Zukunft weitergehende Auswertungen zu. So könnten beispielsweise besondere Einrückungen (etwa zum Erkennen der Lemmata in den Registern), die im ersten Durchgang nicht weiter beachtet werden, auch in Zukunft nach Abschluss des Projektes bei Bedarf relativ einfach berücksichtigt und ausgewertet werden.

Fast schon selbstverständlich ist es, dass bei einem derartigen Projekt sämtliche Daten in standardisierten, langfristig les- und archivierbaren Formaten vorliegen. Die Text- und die Koordinatendaten werden in XML erfasst, das für die Langzeitarchivierung in TEI-XML transformiert wird. Die Bilddateien liegen im TIF-Format vor. Das Vorliegen der Koordinatendaten hat aber auch Auswirkungen auf die Präsentationssoftware mit Volltextsuche. Die MGH sind dem Prinzip der historisch-kritischen Textwiedergabe verpflichtet und haben daher ein erhebliches Spektrum von Textformen abzudecken. Folglich ist es unerlässlich, das historisch gewachsene Erscheinungsbild auch über die bloße Zitierfähigkeit hinaus abzubilden. Daher steht bei der Präsentation immer das Image im Vordergrund – auch bei der Volltextsuche.

Die Suche kann mit der von gängigen Suchmaschinen bekannten Vorgehensweise über die Hauptelemente Editionstext, kritischer Apparat, Sachanmerkungen und sonstiger Text (Einleitungen, Register) durchgeführt werden (vgl. Abb 2). Das Ergebnis der Suche ist jeweils eine Trefferliste, die unter Angabe der Fundstelle zu einem spezifischen Treffer führt. Die Hinzuziehung der Koordinatendaten ermöglichen es, jedes gefundene Wort im Image optisch hervorzuheben („Highlighting“).

Bei Bedarf kann auch die Ansicht des Volltextes für die betreffende Seite aufgerufen werden. Unser gesamtes Korpus an Texten wird im Rahmen der Suchsoftware zunächst einheitlich behandelt werden. Ein Zugriff auf Informationen, die von ihrem Charakter her nur in bestimmten Quellengattungen auftreten, wird zunächst nicht möglich sein. So wird es zum Beispiel keine Möglichkeit geben, in einer Suchanfrage sämtliche Urkunden auflisten zu lassen, die beispielsweise in Regensburg im 9. Jahrhundert ausgestellt wurden. Die zur Verfügung gestellten Abfragemöglichkeiten orientieren sich an dem kleinsten gemeinsamen Nenner, der für unsere Texte zu finden ist.

III. Verlinkung

Digitalisierungsprojekte wie die dMGH müssen nicht nur Rechenschaft darüber ablegen, nach welchen Kriterien das Material ausgewählt wird und wie dieses verarbeitet und präsentiert wird. Darüber hinaus gewinnt in zunehmendem Maße die Frage, wie eine bestimmte Webanwendung mit anderen verwandten Projekten interagiert, an Relevanz. Neben der weiter planmäßig fortschreitenden Texterfassung gilt diesem Problemfeld die größte Aufmerksamkeit, was weitere Entwicklungen und Optimierungen der dMGH angeht. Aufgrund des enthaltenen Materials sind die dMGH sicherlich in erster Linie Endpunkt für Verlinkungen. Geschichtswissenschaftlich orientierte Internetangebote zitieren häufig in den MGH edierte Quellen. Um diese Zitate leichter nachprüfbar zu machen und den über das jeweilige Zitat hinausgehenden Kontext zu verdeutlichen, muss es möglich sein, direkte, langfristig stabile Links auf einzelne Seiten der dMGH zu erzeugen. Dies ist in der Präsentationssoftware vorgesehen. Allerdings enthalten die Links derzeit noch projektinterne Elemente (Datenbank-IDs und Imagenummern). Das ist für ein vernünftiges, effektives Arbeiten freilich nicht ausreichend, erfordert es doch von den verlinkenden Projekten detaillierte Kenntnisse der Datenstruktur innerhalb der dMGH. Ein System, das eine kanonische Verlinkung unter Angabe eines Kurztitels und der Seitenzahl (oder beispielsweise im Falle von Urkunden auch unter Angabe der Urkundennummer) ermöglicht, befindet sich derzeit in Vorbereitung.

Ebenfalls noch in der Entwicklung befinden sich webbasierte Programmierschnittstellen (sogenannte Web-Services), die bestimmte, spezialisierte Anfragen, die nicht direkt über die Suchfunktion durchgeführt werden müssen, erlauben. Der Vorteil solcher Dienste liegt darin, dass sie ein rein maschinelles Kommunizieren der Anwendungen untereinander erlauben, ohne dass manuell Anfragen eingegeben werden müssen.

IV. Zukunftsperspektiven

Wagt man einen weiteren Blick in die Zukunft, so wird man natürlich auch die umgekehrte Notwendigkeit sehen: Die dMGH müssen auch von den jeweiligen digitalisierten Editionsseiten kontextsensitiv auf andere, neuere Forschungen und Projekte verweisen können. Als einfachste Ausbaustufe wäre zum Beispiel ein projektimmanenter Hinweis denkbar, wenn zu einer angezeigten Seite Addenda oder Corrigenda im selben Band vorhanden sind. Ein weiterer Schritt, der über den Rahmen der dMGH hinausweist, wird die enge Verknüpfung der dMGH mit der von Theo Kölzer angelegten Datenbank mit Nachträgen und Ergänzungen zu den von Theodor Sickel 1879-1893 edierten Ottonen-Urkunden sein.² Hierbei handelt es sich um eine reine Online-Publikation, die vor allem Literaturnachträge, aber auch sonstige Ergänzungen oder in Einzelfällen auch Neueditionen zu den Ottonen-Urkunden enthält. Die wissenschaftliche Arbeit mit den Sickelschen Editionen wird zukünftig immer des Rückgriffs auf die Ottonen-Datenbank bedürfen. Zu jeder in den dMGH ange-

zeigten Urkunde sollte der Nutzer per direktem Link auf vorhandene Nachträge aufmerksam gemacht werden.

Aus Nutzersicht wäre es freilich auch wünschenswert, wenn zukünftig andere Forschungsprojekte, die neueres Material zu den MGH-Editionen bieten, verlinkt würden. Hier stellen sich allerdings rein praktische Fragen der Pflege, der Weiterentwicklung und der redaktionellen Überprüfung solcher Verknüpfungen. In diesem weiten, zum Glück für den wissenschaftlichen Fortschritt letztlich unbegrenzten Feld ist sicherlich das letzte Wort noch nicht gesprochen.

Clemens Radl

Zum Autor

Clemens Radl, wissenschaftlicher Angestellter bei den Monumenta Germaniae Historica
und Bearbeiter der digitalen MGH
Kontakt: clemens.radl@mgh.de
Website der dMGH: <http://www.dmgh.de>

Anmerkungen

- 1 Zugänglich ist das Projekt unter <http://www.dmgh.de>. Dort ist auch eine Testversion der Software, die die Volltextsuche ermöglicht, verlinkt.
- 2 Theo Kölzer, Ergänzungen zu den MGH Diplomata regum et imperatorum Germaniae I–II (Urkunden Konrads I. bis Ottos III., 911–1002), <http://www.mgh.de/diplomata/nachtraege.htm>

Auszug aus:

**Jahrbuch der historischen Forschung
in der Bundesrepublik Deutschland 2006**

Herausgegeben von der
Arbeitsgemeinschaft historischer Forschungseinrichtungen
in der Bundesrepublik Deutschland e.V.

© 2007 Oldenbourg Wissenschaftsverlag GmbH, München
oldenbourg.de

Das Werk einschließlich aller Abbildungen ist urheberrechtlich geschützt.
Jede Verwertung außerhalb der Grenzen des Urheberrechtsgesetzes ist ohne
Zustimmung des Verlages unzulässig und strafbar. Dies gilt insbesondere für
Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung
und Bearbeitung in elektronischen Systemen.