Europäisches Urkundenerbe

Zu Potentialen und Perspektiven eines internationalen Fachinformationssystems digitaler Urkundenpublikationen

Georg Vogeler, Ludwig-Maximilians-Universität München

Ausgearbeitete Fassung des Vortrags auf der des „Elektronische Fachinformationssysteme in der Geschichte“ der Arbeitsgemeinschaft Geschichte und EDV e.V., München 25./26. November 2004, Stand August 2005

1 Fachinformation „Historische Quellen“

Selbst der Boom der Zeitgeschichte macht eines nicht obsolet: Das Forschungsobjekt der Geschichtswissenschaft ist Vergangenheit. Da die Historiker natürlich in der Gegenwart leben, müssen sie die Vergangenheit in die Gegenwart holen. Unserer Tagung lieferte dafür das eine oder andere Beispiel. Hier geht es nun darum, das einzige, was wir aus der Vergangenheit haben, unsere Quellen nämlich, mit den Fragestellungen der Zukunft in Verbindung zu bringen. Wie gut sind diese zentralen Objekte historischer Arbeit von elektronischen Fachinformationssystemen erschlossen? Wenn man die Rolle der Quellen im historischen Erkenntnisprozeß ernst nimmt, muß man diese Frage etwas genauer fassen. Es ist dann weniger zu fragen: Welche Angebote von Quellen sind im Internet vorhanden und wie sind sie in den vorhandenen Portalen verzeichnet? als: Wie könnte eine Suchumgebung aussehen, die über Katalogfunktionen hinausgeht und statt dessen Probleme historischer Fragestellungen aufnimmt?

Vielleicht gibt es eine solche Suchumgebung ja schon: Wenn man sich wichtige Portale zur historischen Fachinformation ansieht, dann sind die historischen Quellen jedoch nicht leicht zu finden: Clio Online[1] kennt Foren, Institutionen, Themenportale, Publikationen, einen Stellenmarkt und immerhin im Webverzeichnis eine Kategorie „Materialien“ und einen „Guide“ zur Archivbenutzung. Unter den Institutionen und Portalen gibt es natürlich auch Archive. Die derzeit nur mit dem Internet Explorer benutzbare Metasuche bietet auch die Möglichkeit nach Archivalien und Quellen zu suchen: Die Onlinefindbücher der Archive in NRW, die British Library Map Collection, die IKAR Altkartendatenbank, Pictura Paedagogica Online und das VD17 sind darunter zusammengefaßt. Damit kommt man in den nordrhein-westfälischen Archiven bis auf Bestandsebene, in den übrigen Angeboten bis zu den Metadaten der Bücher, Karten oder Bilder.

[1] http://www.clio-online.de/, letzte Einsichtnahme 3.8.2005

Im Münchener Pendant „Chronicon“ sind umfangreiche Literaturdatenbanken zusammengefaßt, eine Rubrik Webressourcen, ein paar biographische Hilfsmittel und Digitalisate. Damit zeigt Chronicon seine stark bibliothekarische Ausrichtung. Man kann Metadaten von Quellen finden, wenn diese als gedrucktes Buch oder als Webressource vorliegen. Interessant ist, daß mittelalterliche Handschriften, deren Kataloge umfangreich digitalisiert sind,[2] nicht mit ins Angebot einbezogen sind.

Welche Angebote müßten denn in ein quellenorientiertes Fachinformationssystem eingebunden werden? Ich möchte mich beispielhaft auf das Mittelalter konzentrieren. Im Internet finden wir als Quellen zur Geschichte Europas zwischen dem Untergang des Römischen Reiches und der Entdeckung Amerikas umfangreiche Texteditionen unterschiedlichster Themen: Als Beispiele seien genannt: die Neuedition der Sammlung des Benedictus Levita, eines zentralen kirchenhistorischen Textes,[3] Quellensammlungen zur Geschichte der Juden[4] oder die gedruckten Editionen von Geschichtschreibern, die man im Angebot der Bibliothèque Nationale Française „Gallica“ finden kann.[5] Wir haben es dabei mit Abschriften von Originalen zu tun, die mehr oder weniger aufwendig nach den Regeln der Editionswissenschaft zu lesbaren und erklärten Texten umgearbeitet wurden und nun mit Hilfe des Computers gelesen werden können.

Auch die Archive sind reichhaltig im Internet aufzufinden. Wenn man sich Karsten Uhdes Liste der deutschen Archive[6] oder das UNESCO Archives Portal[7] durchsieht, dann findet man jedoch überwiegend nur Adressangaben. Viele Archive haben in den letzten Jahren ihr Angebot mit Kurzbeschreibungen und Beständelisten erweitert. Einige Findmittel sind auch schon präsent, häufig als zentrale Projekte der Archivverwaltungen wie in Baden-Württemberg[8] oder in Nordrhein-Westfalen[9].

Bilder von den Dokumenten findet man selten. Das Stadtarchiv Duderstadt[10] ist zwar nicht mehr das einzige Archiv, das große Teile seiner Bestände digitalisiert hat, aber es darf dennoch hervorgehoben werden, denn es ist ein Pionier auf diesem Gebiet (1997) und große Teilbestände sind

[2] Vgl. die Handschriftendatenbank und die Handschriftenkataloge in Manuscripta Medievalia: http://www.manuscripta-mediaevalia.de/, letzte Einsichtnahme 3.8.2005

[3] Edition der falschen Kapitularien des Benedictus Levita, bearb. v. Prof. Dr. Gerhard SCHMITZ u.a., http://www.uni-tuebingen.de/mittelalter/forsch/benedictus/haupt.htm, letzte Einsichtnahme 4.8.2005.

[4] Quellentexte zur mittelalterlichen Geschichte und Geschichte der Juden, hg. v. Christoph CLUSE, 2003, http://www.uni-trier.de/uni/fb3/geschichte/cluse/texte.htm, letzte Einsichtnahme 4.8.2005

[5] http://gallica.bnf.fr/, letzte Einsichtnahme 4.8.2005.

[6] http://www.archivschule.de/content/59.html, letzte Einsichtnahme 4.8.2005

[7] http://www.unesco.org/webworld/portal_archives, letzte Einsichtnahme 4.8.2005.

[8] http://www.landesarchiv-bw.de/, letzte Einsichtnahme 3.8.2005.

[9] http://www.archive.nrw.de/, letzte Einsichtnahme 3.8.2005.

vollständig online verfügbar. Ähnlich umfangreiche Digitalisierungsprojekte gibt es auch heute nur selten. Originale Quellen sind also im Internet nicht sicher dort zu finden, wo sie aufgehoben werden. Zusätzlich zu den Archivangeboten gibt es Seiten von retrodigitalisierten Abdrucken, von neu erstellen Editionen, von privaten Sammlungen, von Texten in wissenschaftlichen Abhandlungen usw.

2 Urkundensuchmaschine

Wir haben es also mit einer Vielfalt unterschiedlicher elektronischer Repräsentation der Quellen zu tun: Volltexte und Bilder der Quellen, Bilder von gedruckten Quelleneditionen, Metadaten einzelner Quellen, Metadaten von Quellencorpora und Verwahrinstitutionen. Der Ausgangsfrage nach den Suchmöglichkeiten jenseits der Katalogisierung, möchte ich die beiden letzteren Fälle übergehen, da sie den Informationsgehalt der Dokumente nur sehr ungenügend abbilden. Statt dessen möchte ich von den konkreten Überlegungen ausgehen, die im Rahmen eines Projekts an der Ludwig-Maximilians-Universität in Zusammenarbeit zwischen dem Historischen Seminar und dem Institut für Computerlinguistik angestellt werden. Unter dem Arbeitstitel „DING“ („Das Ist Nicht Google“)[11] sollen Möglichkeiten einer fachspezifischen Suchmaschine für mittelalterliche und frühneuzeitliche Urkunden ausgelotet werden. Urkunden bieten sich als Testfall für eine solche Suchmaschine aus drei Gründen an:

1. Urkunden sind eine zentrale Quelle für die Geschichte des Mittelalters.
2. Urkunden sind im Verhältnis zu anderen Quellengattungen relativ gleichförmig und strukturiert.
3. Urkunden werden nach weitgehend anerkannten gemeinsamen Verfahren beschrieben.

Zum ersten Punkt: Urkunden sind zentrale Quellen für die Geschichte des Mittelalters. Der hohe Quellenwert beruht darauf, dass Urkunden häufig die einzigen Zeugnisse historischer Ereignisse sind, und dass sie zusätzlich als Überrestquelle eine erhöhte Glaubwürdigkeit besitzen. Sie sind aber nicht nur Einzeldokumente historischer Ereignisse, die mit dem Fälschungsnachweis der Urkunde aus der Geschichte verschwinden können, sondern auch als Massenprodukt Quelle für historische Erkenntnis. Aus Testamenten etwa können die Historiker nicht nur Rückschlüsse auf die Rechtspraxis des 14. Jahrhunderts ziehen, sondern z.B. auch auf das Verhältnis zum Tod.[12]

---

[10] http://www.archive.geschichte.mpg.de/duderstadt/dud.htm, letzte Einsichtnahme 3.8.2005.
[11] http://www.cis.uni-muenchen.de/~heller/Classes/ding/index.html, letzte Einsichtnahme 4.8.2005.
[12] Vgl. z.B. Paul BAUR: Testament und Bürgerschaft. Alltagsleben und Sachkultur im spätmittelalterlichen Konstanz, Sigmaringen 1989 (Konstanzer Geschichts- und Rechtsquellen N.F. 31); Kerstin DRONSKE: Lübecker Testamente als Quelle zur Kulturgeschichte des Spätmittelalters, in: Beiträge zur hansischen Kultur-, Verfassungs- und Schiffahrts-

Reihenuntersuchungen von spätmittelalterlichen oder frühneuzeitlichen Privaturkunden sind ein Forschungsfeld mit schwerer Krume. Eine Suchmaschine sollte das Pflügen erleichtern.
Zum zweiten Punkt: Urkunden sind im Verhältnis zu anderen Quellengattungen relativ gleichförmig und strukturiert. Als Zeugnisse von Rechtshandeln haben schon die Zeitgenossen darauf geachtet, ihre Urkunden wiedererkennbar zu gestalten, um so Rechtsstabilität zu erreichen. Auch wenn nicht alle Urkunden die formale Beharrlichkeit der Papstkanzlei sehen lassen, so ist der Wandel doch nur selten revolutionär. Anders als bei längeren Texten, die sich zusätzlich durch Abschriften, Redaktionen und Ergänzungen verändern, ist die Informationseinheit „Urkunde" relativ einfach abzugrenzen, indem auch bei den Zeitgenossen nur das Authenticum, d.h. das beglaubigte Original, eigentliche Geltung besessen hat. Das einzelne Dokument ist gewöhnlich datiert und mit der Angabe des Ausstellungsortes geographisch eingeordnet. Das Urkundenwesen kennt einige sehr klare Verantwortlichkeiten von Personen, die an der Erstellung der Urkunde beteiligt waren: Der Aussteller gibt seinen Willen kund, die Zeugen bezeugen den Rechtsakt, Notare und Schreiber sind genannt oder mit Hilfe von diplomatischen Forschungen ermittelbar. Es erscheint also vertretbar, Urkunden digital so abzulegen, daß sie in ihrem Formalisierungsgrad einem Bibliothekskatalog nahe kommen.

2.1 Zum zweiten Punkt: Urkunden werden nach weitgehend anerkannten gemeinsamen Verfahren beschrieben.

Urkunden, obwohl ihre Inhalte von einfachen Privatangelegenheiten bis zu Staatssachen reichen, werden von den Diplomatikern nach relativ einheitlichen Verfahren beschrieben. Jüngst hatte ich ein georgisches Urkundenbuch in der Hand,[13] das ich nicht lesen konnte, aber richtige Vermutungen darüber anstellen konnte, welcher Art von Informationen ich wo auf der Seite finden konnte: Nummerierungen, einleitende inhaltliche Kurzbeschreibungen, Hinweise zur Überlieferung, der eigentliche Text mit Erläuterungen waren identifizierbare Bestandteile dieser Urkundenedition ebenso wie in einem italienischen, französischen oder niederländischen Urkundenbuch. Die Diplomatik ist eine Wissenschaft, die in den über 300 Jahren ihrer Entwicklung eine relativ stabile gemeinsame Vorstellung der Formen, Typen und Herstellungsverfahren von Urkunden gewonnen

---

geschichte, Weimar 1998 (Hansische Studien, 10; Abhandlungen zur Handels- und Sozialgeschichte 31), S. 61-66; José-Ramón JULIÁ VIÑAMATA: Las actitudes mentales de los barceloneses del primer tercio del siglo XIV, in: Anuario de Estudios Medievales 20 (1990), S. 15-51; Lisane LAVANCHY: Ecrire sa mort, décrire sa vie. Testaments de laics lausannois (1400-1450), Lausanne 2002 (Cahiers lausannois d'histoire médiévale 32).

[13] K'art'uli istoriuli sabut'ebis korpusi, Bd. 1: K'art'uli istoriuli sabut'ebi IX-XIII ss., T'bilisi 1984 (Sak'art'velos istoriis cqaroebi 30).

hat, eine Vorstellung, die sich in den 1990er Jahren in einem internationalen Vokabular manifestierte.[14]

Damit kann sich eine Urkundensuchmaschine von sehr allgemeingehaltenen Erschließungskonzeptionen abgrenzen. Panos CONSTANTOPOULOS, Martin DÖRR, Maria THEODORIDOU u. Manolis TZOBANAKIS[15] haben beobachtet, daß historische Dokumente zwei Erschließungsperspektiven haben, nämlich eine objektorientierte oder archivische und eine inhaltsorientierte oder historische. Ihr Vorschlag, die Schlagwörter der Einzelstücke in Facetten zusammenzufassen, bleibt aber bei allgemeinen dokumentarischen Kategorien stecken: Die Leitfragen „Wer, wann, was, wo und wie?“ für die Kategorien Personen und Organisationen, Aktivitäten und Aktionen, Orte, Zeit sowie Objekte ermöglichen zwar grob funktional differenzierte Schlagwortabfragen, können aber nicht die bei Urkunden wie erwähnt relativ klaren funktionalen Beziehungen möglicher Schlagwörter zueinander abbilden: Die Personenkategorie würde so Aussteller, Empfänger und Zeugen zusammen sortieren, obwohl sie grundsätzlich unterschiedlich zum Rechtsinhalt der Urkunde stehen. Es ist zwar noch zu prüfen, ob diese allgemeinen Kategorien für eine entsprechend allgemein gehaltene Quellensuchmaschine von Relevanz sein können, für Urkunden können sie getrost durch die etablierten Konzepte der Diplomatik ersetzt werden.

Bei aller Einheitlichkeit diplomatischer Arbeit sind doch die Angebote im Internet erstaunlich verschiedenartig: Dabei geht es gar nicht so sehr um die unterschiedlichen diplomatischen Formen oder Unterschiede in der wissenschaftlichen Herangehensweise der Bearbeiter als um unterschiedliche technische Konzepte. Es sind nämlich vorrangig drei Wege, auf denen Urkunden ins Internet geraten: Die Retrodigitalisierung, die Archiverschließung und die wissenschaftliche Neuedition. Dabei entstehen zunächst sehr unterschiedlich strukturierte Einheiten: Die Bibliotheken, die ihre alten Bücherbestände digitalisieren, digitalisieren die Urkunden als Teil eines gedruckten Buches. Die Archivare dagegen haben ein Corpus von Urkundenoriginalen vor sich, das sie in die Hierarchie ihres Archivs einordnen und in Provenienzgruppen aufgeteilt verzeichnen. Die einzelnen Urkunden sind also Teil eines Findbuches. Ein wissenschaftlicher Editor geht schließlich meist von einem Corpus aus, den er mit inhaltlichen Gründen abgrenzt: Die Urkunden, die die Stadt Landshut betreffen, werden da ebenso ediert wie die Urkunden, die Kaiser

[14] Vocabulaire international de la diplomatique, hg. v. Maria Milagros CÁRCEL ORTÍ, 2. verbesserte Auflage, València 1997 (Collecció Oberta).

[15] Panos CONSTANTOPOULOS, Martin DÖRR, Maria THEODORIDOU u. Manolis TZOBANAKIS: Historical documents as monuments and as sources, in: Computer Applications and Quantitative Methods in Archaeology Conference, CAA2002, 2-6 April, 2002, Heraklion, Greece (http://www.ics.forth.gr/isl/publications/paperlink/caa2002.pdf, 2002), letzte Einsichtnahme 4.8.2005.

Friedrich II. ausgestellt hat. Die einzelnen Urkunden sind dabei Teil eines Fließtextes „Urkundenedition“ mit gemeinsamer Einleitung und Registern. Eine Suchmaschine müßte sich von all diesen Einheiten lösen. Anders als bei Google wären nicht die Bildseiten eines Buches oder die HTML-Seite mit einer Liste von Urkunden von Interesse. Die einzelne Urkunde wäre das zu verzeichnende Objekt, egal in welcher zufälligen elektronischen Form und in welcher logischen Hierarchie sie vorliegen.

Um eine solche Aufbereitung schon bei der Erstellung von digitalen Urkundenrepräsentationen vorzubereiten, hat sich im Frühjahr 2004 eine internationale Arbeitsgruppe gebildet, welche die verschiedenen Zugänge in einem gemeinsamen Erschließungsstandard mit Hilfe von XML integrieren will. Die Charters Encoding Initiative (CEI [16]) sieht dabei keine absoluten technischen Formate vor, sondern versucht eine Semantik zu entwickeln, die in unterschiedlichen Strukturkontexten die einzelne Urkunde und ihre zentralen Merkmale doch identifizierbar macht. Sie arbeitet deshalb an Namen für XML-Elemente, die spezifische Phänomene der Urkundenerschließung bezeichnen und mit Hilfe des erwähnten Vokabulars[17] beschrieben werden. Die CEI will ihre Vorschläge in die Konzepte der Text Encoding Initiative[18] (TEI) integrieren, mit deren Hilfe schon gute digitale Editionen entstanden sind,[19] ohne sich darauf jedoch festzulegen: Der Archivar soll damit ebenso Urkunden beschreiben können wie ein Forscher, der sich mit gedruckten Urkundeneditionen auseinander setzt. Noch kann man nicht davon ausgehen, daß alle Angebote einem solchen Standard folgen. Auf die Arbeit der CEI ist bei der Konzeption einer Suchmaschine aber dennoch zurückzukommen.

2.2 Urkunden im Internet

Vorher soll kurz das Spektrum der Angebote, die von der Suchmaschine erfaßt würden, vorgestellt werden:[20] Das Verzeichnis diplomatischer Internetangebote, das in der Virtual Library Geschichtliche Hilfswissenschaften[21] zusammengestellt ist, ist umfangreich. Es verzeichnet z.Zt. 183 Angebote mit sicher mehr als 50.000 Urkunden, und davon sind nur wenige Abhandlungen

[16] http://www.cei.lmu.de, letzte Einsichtnahme 4.8.2005.
[17] Vocabulaire international de la diplomatique, hg. v. Maria Milagros CARCEL ORTI, 2. verbesserte Auflage, València 1997 (Collecció Oberta).
[18] http://www.tei-c.org, letzte Einsichtnahme 27.7.2005.
[19] Vgl. insbesondere die Angebote der Ecole Nationale des Chartes: http://elec.enc.sorbonne.de, letzte Einsichtnahme 3.8.2005.
[20] Eine besser strukturierte Bestandsaufnahme ist der Beitrag Patrick SAHLE u. Georg VOGELER: Urkundenforschung und Urkundenedition im digitalen Zeitalter, in: Geschichte und Neue Medien. Kongreß .hist2003, Berlin April 2003, Berlin 2005 (Historisches Forum. Schriftenreihe von Clio-online 5), im Druck.
[21] http://www.vl-ghw.lmu.de/diplomatik.html, letzte Einsichtnahme 4.8.2005.

über Urkunden, die Mehrzahl enthält Zusammenfassungen, Texte und Bilder von Urkunden. Einzelne Angebote sind auch systematisch: Da sind zunächst einzelne Archive, die ihre Urkundenbestände vollständig über das Internet erschlossen haben: z.B. das Stadtarchiv Duderstadt, das Stadtarchiv Passau oder das Archivio di Stato Roma.[22] Ebenso wie diese Angebote liefern auch die Kaiserurkunden in Abbildungen Bilder der Urkunden, wenn auch in der systematischen Auswahl, mit der Theodor v. Sickel und Heinrich v. Sybel 1888-1891 die Entwicklung der Kaiser- und Königsurkunden des Reiches illustrieren wollten.[23]

Dagegen sind die Urkunden auf ihre Inhalte verkürzt in einer zweite Gruppe: Als Beispiel für Online-Regesten ist an erster Stelle nicht nur wegen der hohen Relevanz des Materials sondern auch weil es umfangreich historische Informationen in die digitale Welt überführt, das Angebot der Regesta Imperii zu nennen.[24] Die Funktionalitäten von Online-Regesten demonstrieren auch sehr gut die hessischen Regesten, in denen 8.756 Urkunden durchsuchbar sind.[25]

Es sind aber auch Volltexteditionen von Urkunden im Internet zu finden: Der Codex Diplomaticus Saxoniae Regiae[26] gehört dazu ebenso wie die dMGH[27] und – als vielleicht technisch anspruchsvollstes Projekt – das regionale Urkundenbuch der Lombardei, das direkt aus den Quellen geschöpft ins Internet gespeist wird.[28] Als letztes Beispiel sei noch eine Datensammlung erwähnt, die mehr als nur regestenartiger Nachweis, Bild des Originals oder Edition ist: Die Anglo-Saxon-Charters bieten ein Netzwerk an Informationen rund um die Urkunden der angelsächsischen Könige.[29]

Damit ergibt sich ein vierter Grund, warum mittelalterliche Urkunden ein geeignetes Testobjekt für eine Quellensuchmaschine sind: Es ist schwer, ähnlich umfangreich und systematisch andere Quellengattungen im Netz zu finden. Aus der Neuzeit sind zwar einzelne Nachlässe oder Korrespondenzcopora online, verschiedene erzählende Quellen des Mittelalters findet man auch auf CD. Aber so systematisch wie die Regesta Imperii, die Quellenaussagen zur Reichsgeschichte vom 9. Jahrhundert bis in 13. und danach in stetig wachsendem Maße umfassen, oder die dMGH,

[22] Duderstadt: http://www.archive.geschichte.mpg.de/duderstadt/dud.htm, letzte Einsichtnahme 3.8.2005; Passau: http://www.stadtarchiv-passau.de/stadtarchiv/bestaende/archbest/urkunden.htm, letzte Einsichtnahme 16.8.2006, die auf ASP basierende Seite ist z.Zt. mit den Mozilla Browsern nicht einsehbar; Rom: http://www.asrm.archivi.beniculturali.it/, letzte Einsichtnahme 16.8.2005.
[23] http://mdz.bib-bvb.de/digbib/urkunden1/kuia/, letzte Einsichtnahme 16.8.2005.
[24] http://www.regesta-imperii.org, letzte Einsichtnahme 4.8.2005.
[25] http://online-media.uni-marburg.de/ma_geschichte/lgr/, letzte Einsichtnahme 4.8.2005.
[26] http://isgv.serveftp.org/codex/, letzte Einsichtnahme 4.8.2005
[27] http://www.dmgh.de, letzte Einsichtnahme 3.8.2005.
[28] Codice diplomatico della lombardia medievale (CDLM): http://cdlm.unipv.it, letzte Einsichtnahme 3.8.2005.
[29] http://www.trin.cam.ac.uk/chartwww/charthome.html, letzte Einsichtnahme 16.8.2005.

in denen die Urkunden der deutschen Herrscher von den Merowingern bis zu Friedrich Barbarossa[30] allen Regeln der diplomatischen Editionskunst folgend im Netz stehen, ist kein anderes Quellencorpus.

3 Nutzen

Es scheint absehbar, daß die Urkunden als beinahe vollständiges europäisches Corpus online vorliegen werden. Diese Entwicklung wirkt auf die Wissenschaft zurück. Von der Rolle der Urkunden als Quelle für historische und sprachhistorische Fragestellungen einmal abgesehen, möchte ich hier aus der Perspektive des Diplomatikers einige konkrete Forschungsperspektiven eröffnen, die mit einem solchen europaweiten Urkundencorpus mit Hilfe einer Suchmaschine untersucht werden könnten. Die Perspektiven folgen der schon 1988 formulierten Erkenntnis von Carlrichard BRÜHL, daß der Vergleich großer Mengen von Urkunden der entscheidende Schritt für die Diplomatik war.[31] Forschungsansätze stellen natürlich auch Forderungen an die Suchmaschine:

3.1 Formularfragen

Zunächst erleichtert der Computer natürlich statistische Auswertungen. Die Aussagekraft einer Statistik der Urkundensprache ist noch umstritten: Benoît-Michel TOCK ist eher skeptisch, weil sein Eindruck aus dem ca. 5000 Urkunden umfassenden Corpus des ARTEM ist, daß nur einige juristische Wörter zeittypisch sind, der größte Teil des Wortschatzes aber nicht.[32] Michel PARISSE war da mit einem sehr kleinen Corpus eher optimistisch,[33] und das Vorgehen des DEEDS-Projekt mag ihm Recht geben, indem es in einem ca. 7000 Urkunden umfassenden Corpus Wortmuster ausweist, die nur in bestimmten Zeiten vorkommen.[34] Solche Studien sollten mit Hilfe der Suchmaschine vermehrt möglich sein.

---

[30] Mit der hoffentlich bald geschlossenen Lücke der Urkunden Kaiser Ludwigs des Frommen.

[31] Carlrichard BRÜHL: Die Entwicklung der diplomatischen Methode im Zusammenhang mit dem Erkennen von Fälschungen, in: Fälschungen im Mittelalter, Bd. 3, Hannover 1988 (Monumenta Germaniae Historica. Schriften 33,3), S. 11-27.

[32] La base de donées des chartes originales antérieures à 1121 conservées en France, in: Resourcing sources, hg. v. Katharine S. B. KEATS-ROHAN, Oxford 2002 (Prosopographica et genealogica 7), S. 153-163.

[33] A propos du traitement automatique des chartes. Chronologie du vocabulaire et repérage des actes suspects, in: La lexicographie du latin médiéval et ses rapports avec les recherches actuelles sur la civilisation du Moyen-Age, Paris, 18-21 octobre 1978 (Colloques internationaux du C.N.R.S., 589) , Paris 1981, S. 241-249.

[34] Die Publikationen von Michael GERVERS und seinem Team, in denen die Methoden und ihre Ergebnisse gezeigt werden, sind umfangreich. Hier sei deshalb nur die Homepage des Projektes (http://www.utoronto.ca/deeds/research/research.html), einen Sammelband (Dating undated Medieval Charters, hg. v. Michael GERVERS, Kongress Budapest 1999, Woodbridge 2000) und den jüngsten Bericht (Michael GERVERS u. Michael MARGOLIN: Application of Computerized Analyses in Dating Procedures for Medieval Charters, in: Le médiéviste et l'ordinateur 42 (http://lemo.irht.cnrs.fr/42/mo42_01.htm, 2003), letzte Einsichtnahme 3.8.2005.) verwiesen.

Die Arbeit des Urkundenforscher ist zur Zeit eher am Aussteller orientiert: Die feinere diplomatische Analyse fragt immer nach der Organisation der Ausstellerkanzlei und versucht sie aus den Urkunden eines Ausstellers zu ermitteln. Dabei ist der Einsatz von Computern nur bedingt effektiv, wie Nicholas BROUSSEAU[35] am Beispiel der vorhandenen Edition der Urkunden Ludwigs des Deutschen plausibel gemacht hat, in denen die Wortstatistik keine signifikanten Abweichungen zwischen Fälschungen und Originalen liefert. Die größten, statistisch nachweisbaren Abweichungen sind nämlich auf Empfängereinfluß und Vorurkunden zurückzuführen. Brousseau reagiert auf dieses Scheitern wie andere Diplomatiker auch schon: Diktatuntersuchungen haben immer das Konzept der Empfängerausfertigung parat, wenn einzelne Urkunden eines Ausstellers zu stark von einander abweichen. Empfängerausfertigungen fallen jedoch nur im Kontrast zu den Kanzleiausfertigungen auf. Eine Suchmaschine, welche die überlieferten Urkunden als Gesamtcorpus behandeln kann, braucht für Formularuntersuchungen keine Unterscheidung zwischen Kanzlei- und Empfängerausfertigung mehr. Sie kann Diktatuntersuchungen über ein nicht mehr vom Aussteller, sondern nur zeitlich und regional begrenztes Corpus vornehmen und so der Frage nachgehen, ob sich sprachliche Spuren eines Notars über die Grenzen der Kanzleien hinweg finden lassen.

3.2 Corpusbildung über Metadaten

Eine weitere Grenze der bisherigen Arbeitsweise kann durch die Integration einzelner Projekte in eine gemeinsame Suchmaschine überwunden werden: Der Forscher kann ein Corpus nach seinen Interessen zusammenstellen. So wie die Regesten Kaiser Friedrichs III. in gedruckter Form bislang nur nach Lagerort geordnet sind, können die Forscher auf der CD bzw. inzwischen in der Onlinefassung[36] der ursprünglich in weiter Zukunft vorgesehenen Gesamtfassung vorgreifen und eine chronologische Gesamtreihe erstellen. Mit Hilfe einer projektübergreifenden Suchmaschine könnten also aus städtischen und klösterlichen Urkundenbüchern, die ihre Corpora nach Betreff gebildet haben, neue Corpora nach Aussteller oder nach Empfänger gebildet werden. Jedes mögliche Suchkriterium bildet nämlich aus der Gesamtmenge der Urkunden eine Teilmenge, die für sich weiteren Untersuchungen unterzogen werden kann.

[35] Nicholas BROUSSEAU: Lemmatisation et traitement statistique. De nouveaux instruments pour la critique diplomatique?, Le cas des diplômes pseudo-originaux au nom de Louis le Germanique, in: Médiévales 42 (2002), S. 27-43.
[36] http://www.regesta-imperii.de

3.3 Überregionale Bezüge

Schließlich dokumentiert die Suchmaschine überregionale Zusammenhänge, die bislang in der Vielfalt der lokal, regional oder national organisierten Projekte untergingen. So wären z.B. die Ausbreitung des Notariatsinstruments oder der Siegelurkunde[37] in einem gesamteuropäischen Urkundencorpus, wie er sich in der Suchmaschine wiederfände, geographisch und zeitlich genau verortbar und je nach angewendetem Analysetool auch kartierbar, graphisch darstellbar etc. Ebenso könnte man die Wanderung von Formelgut nachvollziehen. Die Reflexe zwischen der gesamteuropäischen Papsturkunde und lokalen bzw. regionalen Urkundenkulturen würden sichtbar. Teilweise schon alte Fragen der Urkundenlehre würden damit einer Antwort näher kommen.

4 Konzeption der Urkundensuchmaschine

Es ist also zu fragen, was eine wissenschaftliche Internetsuchmaschine für Urkunden leisten müßte, um ein solches virtuelles Corpus den Fragen der Forscher zugänglich zu machen. Dabei kommen die Ergebnisse der Arbeit der CEI zum Tragen: Das Konzept einer urkundenspezifischen Suchmaschine würde sie als eine Art Ontologie zu nutzen versuchen, mit deren Hilfe die Benutzeranfragen ausgewertet und mit der Struktur der Angebote korreliert werden können. So wäre z.B. der Index der Suchmaschine nach dem Modell der CEI aufzubauen. Die Oberfläche der Suchanfrage könnte Elemente einer CEI-Ontologie aufnehmen und damit auf die inhaltlichen Interessen des Suchenden schließen, um die Treffergenauigkeit bewerten zu können. Die Suchmaschine darf aber auf keinen Fall von vorab formulierten Fragen ausgehen, sondern muß folgende Probleme typischer Forschungsmethoden zu lösen helfen.

4.1 Grundkategorien und Zugriffsperspektiven

Wie auch in der Kategorisierung von CONSTANTOPOULOS et al.[38] sind einige Fragen des Zugangs zu historischem Material unterschiedlichen Nutzern gemeinsam: Chronologie und Geographie dienen als Orientierungs- und Eingrenzungsmerkmale und müssen als solche Teil der Suchumgebung sein. Die Philologie der Texte, d.h. insbesondere ihre orthographische und morphologische Struktur ist ein ebensolches zentrales Kriterium. Welche Problem mit diesen Grundkategorien verbunden sind, soll weiter unten erläutert werden.

Zunächst ergeben sich aber essentielle methodische Differenzen zwischen wahrscheinlichen Nutzergruppen einer solchen Suchmaschine: Urkunden sind nämlich nicht nur für Historiker eine

[37] vgl. dazu auch Benoît TOCK, wie Anm. 32.
[38] Vgl. Anm. 15.

wichtige Quelle, sondern sie bieten vielfach auch die ältesten volkssprachlichen Zeugnisse. Philologen und Linguisten fragen aber anders als Historiker: Während für letztere der Wert einer Quelle in ihren sachlichen Inhalten liegt und so z.B. die Orthographie eines Personennamens nur als Indiz für eine Identifikation dient, ist für den Sprachhistoriker der schriftlich dokumentierte Lautbestand gerade das Untersuchungsobjekt.

Es ist also nötig, die Suchumgebung drei unterschiedlichen Fragestrategien anzupassen:

1. Eine historische muß normalisierte Informationen höher bewerten als nicht normalisierte. Darüber hinaus ist externes Wissen nutzbar zu machen: Die Stichwortsuche sollte auf Lemmata verweisen, geographische Informationen sollten hierarchisiert sein und orthographische Varianzen sollten durch Zeichenausgleichsverfahren wie Soundex verringert werden.
2. Eine linguistisch-philologische Zugangsweise kann diesen sprachlichen Ausgleich auch verwenden, muß ihn aber explizieren, d.h. bestimmen können, welche Verfahren anzuwenden sind. Darüber hinaus sind sprachorientierte Vergleichsmuster vorzusehen: Eine „nahebei"-Funktion sollte z.B. einfache grammatische Strukturen wie Sätze als Grenzen berücksichtigen können.
3. Die dritte Perspektive ist eine genuin urkundenwissenschaftliche. Eine diplomatische Suchumgebung würde vom einzelnen Urkundenobjekt ausgehen und formale Kriterien wie „Siegel vorhanden", Format oder den Beschreibstoff ebenso in die Suche mit einbeziehen wie die diplomatische Struktur des Textes. Mit einer derartigen Umgebung könnten statistische Übersichten über das Vorkommen z.B. bestimmter Datierungsstile nach Zeit, Gattung, Aussteller und Region erstellt werden.

Die Suchmaschine muß diese drei Zugriffe insbesondere in das Ranking der Suchergebnisse einbeziehen. Im Suchvorgang sind aber noch weitere Probleme des Corpus zu berücksichtigen: die unterschiedliche Erschließungstiefe der Vorlagen, Unschärfen in der Suchanfrage, der Umgang mit Datumsangaben, linguistische Ausgleichsmechanismen und Verfahren zur Bewertung.

4.2 Ungleich tiefe Strukturierung

Bei aller postulierten Ähnlichkeit innerhalb der Quellengattung und Formalisierung des wissenschaftlichen Umgangs damit bleibt doch immer eine größere Individualität in der Bearbeitung durch einen Wissenschaftler. Wie erwähnt, können Urkunden als Regesten, Faksimiles oder Volltexteditionen repräsentiert werden, können sie als proprietäre Datenbanken, einfache HTML-Texte oder mit Hilfe von XML mit einer Textmetaschicht versehen sein. Die Strukturen der Corpora gehen von unterschiedlichsten Wurzelelementen aus: Retrodigitalisierte Bücher, Empfän-

gerprovenienzen und Ausstellercorpora sind nur die drei wichtigsten. Die Bearbeiter arbeiten mit unterschiedlichen Budgets und unterschiedlichen Ansprüchen an die Erschließungstiefe. Die Suchmaschine kann also nicht wie ein Datenbankinterface eines Bibliothekskatalogs harte Grenzen zwischen den Eigenschaften eines Dokuments ziehen. Sie muß damit umgehen können, daß eine Person einmal als solche im Fließtext hervorgehoben ist und das andere Mal nicht.
Insofern folgt die Suchmaschine traditionellen Konzepten der Suchmaschinentechnologie: Sie durchsucht zunächst immer das ganze Dokument. Nur die Ergebnispräsentation wird inhaltlich gesteuert. Die Suchmaschine muß also die Relevanz einer Fundstelle nach ihrem logischen Kontext bewerten: Eine historische Suchanfrage nach einer Person wird in einem Dokument mit tiefer Erschließung und Normalisierungen genauer sein dürfen als in einem einfachen Volltext.

4.3 Ähnlichkeitsmodell

Fragen an Sprachdokumente und historische Zeugnisse gelten als wenig „exakt". Dieser Eindruck entsteht durch die Vielfalt der Möglichkeiten, mit denen verschriftlichte Sprache Sachinformationen kodieren kann. Die Suchanfragen an ein europaweites Corpus werden also eher heuristischen Konzepten folgen, die in mehreren Stufen fragen. Die Suchmaschine muß deshalb Suchen nach Ähnlichkeiten ermöglichen. Unser Suchmaschinenprojekt arbeitet dabei an einer Umgebung, die ein „Query by Example", ein „Suchen nach Vorbild", ermöglicht. Die Eigenschaften einer Urkunde sollen also in ein Anfrageschema übersetzbar sein, das mit vorgegebenen Toleranzen Abweichungen zuläßt. Dabei sind natürlich zunächst die erwähnten linguistischen Ähnlichkeiten und Nachbarschaften anzuwenden. Darüber hinaus sind formalisierte externe Eigenschaften für Sachinformationen zu verwenden: Eine Person wird neben dem Namen auch über Funktionen und institutionelle Zuordnungen beschrieben. Soweit derartiges Wissen im Corpus expliziert ist, ist es für die Einschätzung der Ähnlichkeiten zu berücksichtigen. Bei geographischen Angaben erscheint es möglich, externe Wissensbasen anzugliedern und damit Einzelorte und Räume einander zuzuordnen und zu hierarchisieren. Eine Suche nach München darf also auch Treffer in Bayern finden. Das erwähnte Schema der CEI berücksichtig darüber hinaus die formalen Klassen, in denen die Urkunden in den Archiven und von den Editoren beschrieben werden. Die Funktionen von Inhalten, die in Elementen wie Aussteller, Empfänger, Ausstellungsort, Zeitraum, Urkundenklasse oder Rechtsinhalt enthalten sind, sind Kriterien, die Ähnlichkeiten beschreiben können: Die Diplomatik hat uns gelehrt, daß der Aussteller ebenso Auftraggeber einer Urkundenproduktionsstelle sein kann, wie der Empfänger, daß neben diesen aber auch der Aus-

stellungsort auf die Form der Urkunde wirken kann, wenn z.B. in der Kanzlei der deutschen Kaiser und Könige auf ihren Romzügen italienische Schreiber beschäftigt worden sind. Zwei Urkunden werden deshalb einander sehr ähnlich sein, wenn sie vom selben Aussteller für den gleichen Empfänger am selben Ort ausgestellt worden sind, weniger ähnlich, wenn sie zwar vom selben Aussteller, aber als ein anderer Urkundentypus ausgefertigt worden sind.

4.4 Chronologie

Es erscheint als Selbstverständlichkeit, Zeitangaben in eine historisch orientierte Suchmaschine einzubinden. Dabei entstehen aber Probleme, die mit einfacher Angabe einer Datierung nicht gelöst sind: An sich sind Datumsangaben in Urkunden auf den Tag genau: *X kal. Junii Anno incarnationis M^{o} C^{o} LVI^{o}*. Der Suchende wird jedoch zunächst häufig genug nicht nach einem einzelnen Datum sondern nach einem Zeitraum fragen: 1154-1156. Es gilt also, Datumsangaben nicht als Text, sondern als numerisch skalierbare Werte zu suchen. Komplexere Probleme entstehen aber besonders dann, wenn die Urkunden im Corpus nicht so genau datiert sind wie oben angegeben. Es ist nicht nur eine Besonderheit der englischen Privaturkunden des 13. Jahrhunderts, daß sie keine Datumsangabe tragen. Dennoch lassen sie sich häufig auf Grund des Inhalts oder der äußeren Merkmale grob datieren: „zwischen 1155 und 1158“, „Sommer 1155“, „vor 1190“, „Mitte 12. Jahrhundert“. Einfach erscheint es zunächst, sprachlichen Datumsangaben („Anfang 12. Jahrhundert“) eindeutige Werte zuzuorden (=1100-01-01 – 1115-12-31). Darüberhinaus muß die Suchmaschine aber auch Bewertungsalgorythmen für Überlappungen von Zeiträumen und Vergleichsmuster für offene Zeitspannen („vor 1190“ = `NULL` – 1189-31-12) anbieten.

4.5 Philologie

Das Suchinterface soll dem Benutzer die Möglichkeit geben, anzugeben, wie nahe seine Suchanfrage am Zeichenbestand der Texte sein soll, ob er Flexionen berücksichtigen will oder nicht, wie er mit dem paläographischen Befund umgehen will, kurz: er wird die Granularität der sprachlichen Repräsentanz seiner Anfrage einstellen wollen. Das Corpus kann nämlich Urkundentexte von der diplomatischen Abschrift unter Berücksichtigung aller Abkürzungen, graphischen Besonderheiten oder Diakritika bis zu modernisierten Fassungen in einem Regest enthalten:
… in villa que dicit<abbr>ur</abbr> <place norm=„Malsbach“>Alersbach</place> …
… in villa, quae dicitur Malsbach ...

Um eine Anfrage hier zielgenau entweder zu individuellen Sprachformen oder zu allen Zeugnissen einer benannten Sache zu führen, muß die Suchmaschine verschiedene Informationen mitberücksichtigen: In den für Urkundendigitalisate verbreiteten Auszeichnungssprachen (TEI, CEI, DTD des Codice Diplomatico della Lombardia Medievale) ist ein Teil der Informationen über Tags dargestellt (hier: `<abbr>`). Solche linguistischen Tags gehören also in die Analyse der Anfrage.

Andere Formen der sprachlichen Repräsentation sind über linguistische Analysen zu ermitteln. Dabei geht es insbesondere um Lemmatisierung und um orthographischen Ausgleich: *aecclesiam* –> *ecclesia*. Für letzteren stehen teilweise auch schon für historisches Material erprobte Verfahren des Zeichenausgleichs zur Verfügung, die insbesondere für Identifizierung von gleichen Namen entwickelt worden sind.[39] Für erstere gibt es natürlich Lexika und morphologische Analyse-Tools, die jedoch einen entscheidenden Nachteil haben: Sie beschränken sich meist auf die moderne Form einer Sprache. Es gälte also, das in den entsprechenden Philologien vorhandene Wissen über die Lautentwicklung und die historischen Änderungen der Morphologie in angemesse Alogrythmen zu gießen, die eine Ähnlichkeitsverwertung erlaubten.

4.6 Bewertungen

Eine erste Aufgabe der Suchmaschine wäre also eine inhaltlichen Analyse des durchsuchten Corpus, um auf unterschiedlichen Abstraktionsebenen suchen zu können. Daneben tritt aber ein zweites Konzept, das aus einer allgemeinen Volltext-Suchmaschine eine fachspezifische Suchmaschine macht, nämlich das nach Frageart und Erschließungstiefe der Teile des Corpus differenzierte Bewertungsschema. So sind orthographische Varianten evtl. linguistische Varianten, aber sehr unwahrscheinlich sachliche Varianten. Es gälte also ÄhnlichkeitsAlogrythmen für den Merkmalsvergleich zu entwickeln, die von der Suchperspektive gesteuert werden – ohne daß der Benutzer umfangreiche Feineinstellungen auf Grund seiner Vorkenntnis von Corpus und Suchmaschine vornehmen muß. Unser Konzept der Suchmaschine wird mit seinen drei Benutzertypen hoffentlich solche Alogrythmen liefern können.

Bei der Bewertung muß aber neben den Ähnlichkeiten auch die Struktur des Textes bzw. den Ort der Informationen in der Struktur berücksichtigen: Überschriften beispielsweise enthalten Informationen, die für einen ganzen Textabschnitt relevant sind, ohne daß in diesem dann noch die

[39] A.J. LAIT u. B. RANDELL: An Assessment of Name Matching Algorithms (http://homepages.cs.ncl.ac.uk/brian.randell/home.informal/Genealogy/NameMathing.pdf, ohne Datum), letzte Einsichtnahme 3.8.2004.

entscheidenden Stichwörter vorkommen müssen.[40] Dabei ist die Auszeichnungstiefe des jeweiligen Teilcorpus ein diese Einschätzung modifizierender Faktor.

5 Fazit

Auch wenn viele Bestandteile der Suchmaschine noch Postulate sind: eine urkundenspezifische Suchmaschine scheint möglich und sinnvoll. Das liegt einerseits am stark formalisierten Material und den stark systematisierten Verfahren der medialen Repräsentation nicht nur im Buchdruck sondern auch in der digitalen Welt. Andererseits ist die Quellengattung „Urkunde“ besonders reichhaltig im Internet vertreten und verspricht interessante Forschungsergebnisse. Eine fachspezifische Suchmaschine, wie wir sie in München für Urkunden konzipieren, würde den geläufigen Forschungen an Urkunden neue Fragen hinzufügen. Sie würde darüber hinaus aber auch Technologien entwickeln, die für die Suche in anderen Quellengattungen nutzbar wären: So würden z.B. Techniken für den Umgang mit Datumsangaben entwickelt. Erkenntnisse diachroner Linguistik wären in Programmmodule zur Lemmatisierung oder zum Zeichenausgleich zu überführen. Es wären allgemeine Alogrythmen zu entwickeln, Textstrukturen in die Bewertung von Treffergenauigkeiten einzubeziehen. Neben linguistischen Ähnlichkeitsmodellen wird auch über quellentypische nachgedacht. Eine allgemeine Quellensuchmaschine bräuchte dann nur noch einen den Diplomatikern vergleichbaren Impetus, umfangreiches Material online zu stellen. Vorerst wäre das „DING“ aber besonders eine virtuelle Repräsentation des europäischen Urkundenerbes – und das ist alleine schon eine spannende Perspektive.

[40] Vgl. dazu auch Norbert FUHR: Information Retrieval Methods for Literary Texts, in: Jahrbuch für Computerphilologie 5 (2003), S. 147-160.