Elektronische Wörterbücher

Vom Buch ins Web

Der Medienwandel bringt uns vom Buch ins Web, vom Zettelkasten zum elektronischen Suchindex. Wenn nicht gerade der Strom ausfällt oder der Internetprovider Probleme macht, haben digitale Wörterbücher große Vorteile. Schritt für Schritt gehen derzeit auch die wissenschaftlichen Wörterbücher der Bayerischen Akademie der Wissenschaften „ins Netz".

Von Alexandra Gobrecht

Jedermann kann überall auf der Welt Webseiten aufrufen und nutzen, wenn er im Besitz eines internetfähigen elektronischen Geräts und einer Netzverbindung ist. Man findet schneller Antworten auf seine Fragen und schätzt die umfangreichen Suchmöglichkeiten, die ein Buch nicht bieten kann, zum Beispiel die Suche in verschiedenen Feldern wie Wörterbuchtext oder Stichwort. Man bekommt oft schon beim Eingeben eines Begriffs mögliche Ergänzungen angezeigt und kann seine Recherche mit Operatoren oder Platzhaltern wie einem Asterisk erweitern. Den Suchbegriff sieht man in den Volltexten farbig hervorgehoben, große Treffermengen kann man nach verschiedenen Kriterien filtern oder sortieren. Von einem Wörterbuchartikel aus führen Links mit einem Klick zu zusätzlichen Informationen wie Angaben aus dem Quellenverzeichnis, weiteren im Text erwähnten Wörtern mit eigenem Eintrag oder anderen Webangeboten, die ebenfalls Informationen zum selben Stichwort enthalten. Umfassender kann eine Recherche in Nachschlagewerken nicht sein.

Ist ein Wörterbuch nun zuerst in Buchform erschienen, so kann es mit verschiedenen Verfahren retrodigitalisiert werden: entweder durch Einscannen der Seiten und anschließende optische Zeichenerkennung, die einen so genannten Volltext generiert, oder durch Abschreiben der Druckseiten, wobei mehrere Datentypisten denselben Text erfassen. Die Genauigkeit lässt sich in beiden Fällen mit Computerprogrammen verbessern, die die Textfassungen abgleichen. Es gibt heute aber auch genuin digitale Wörterbücher, und ihre Zahl steigt stetig an.

DIE AUTORIN

Die Computerlinguistin Alexandra Gobrecht ist seit Sommer 2011 Ansprechpartnerin für alle Fragen der Retrodigitalisierung, digitalen Publikation und Langzeitarchivierung in der Bayerischen Akademie der Wissenschaften.

TEI, ein Standard der Datenbeschreibung für Geistes- und Sozialwissenschaften

Ob Texte nun retrodigitalisiert werden oder von Beginn an digital entstehen, ein Standard der Datenbeschreibung etabliert sich: TEI, ein von der „Text Encoding Initiative" entwickelter XML-Dialekt für die Auszeichnung von Textdokumenten der Geistes- und Sozialwissenschaften. Mit diesem Dokumentenformat können inhaltliche, strukturelle und konzeptuelle Eigenschaften, aber auch Informationen über das Layout direkt im Volltext markiert werden. TEI wird von einem nicht gewinnorientierten Konsortium publiziert, das sich aus Vertretern von Universitäten, Bibliotheken, akademischen Projekten und anderen Forschungseinrichtungen zusammensetzt.

Einträge in Nachschlagewerken sind relativ komprimiert: Sie enthalten sehr viele implizite Informationen auf mehreren Ebenen. Einem geübten Wörterbuchnutzer erschließen sie sich vielleicht auf den ersten Blick, für die maschinelle Verarbeitung von Texten ist es jedoch notwendig, solche Informationen explizit zu machen. Das geschieht in XML durch Markierung des Volltextes mit Auszeichnungen in spitzen Klammern, so genannten Tags. TEI besteht aus einem Kern für allgemeine Auszeichnungen und Modulen für spezielle Textgattungen wie Wörterbücher. TEI ist ein offenes, freies Format, das sich maschinell weiterverarbeiten lässt, zum Beispiel durch

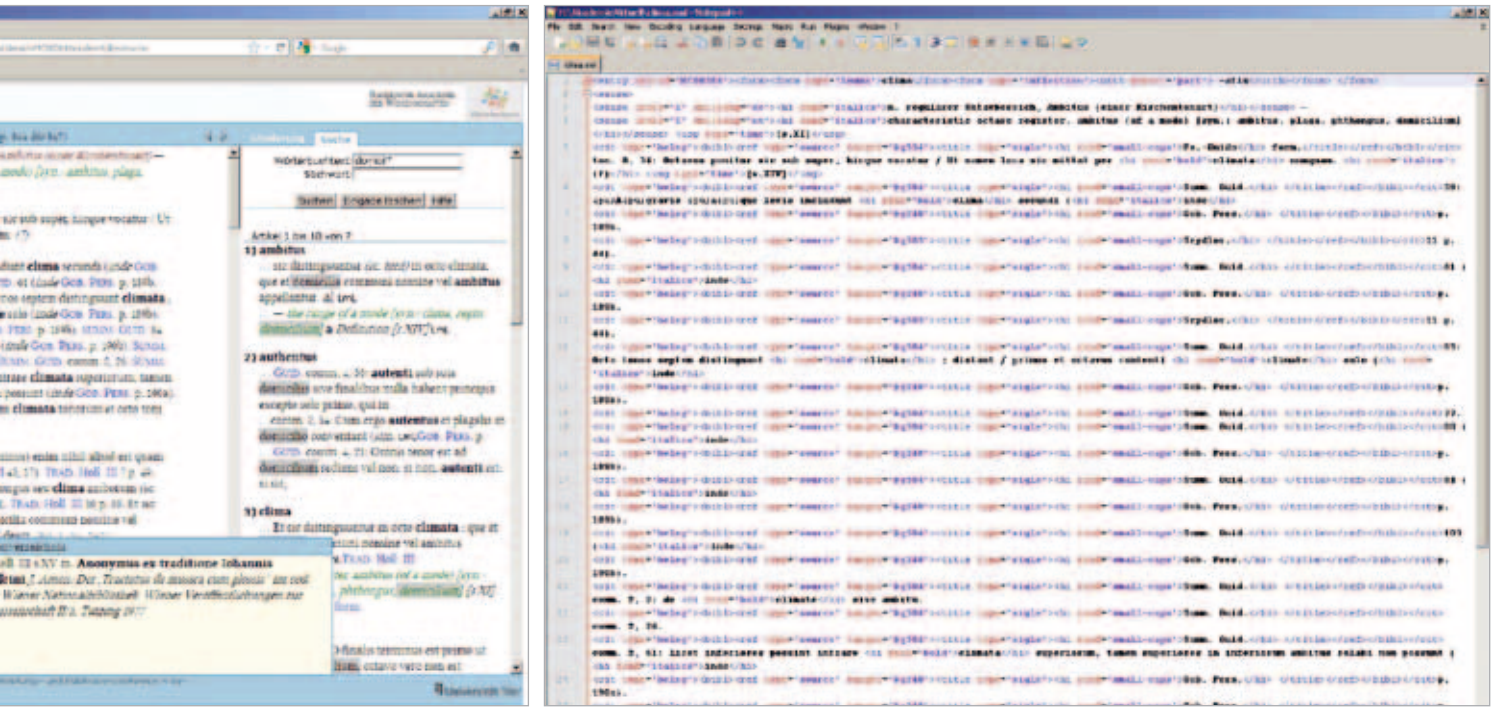

Abb. 1: Die Buchstaben A bis D des Lexicon musicum Latinum sind bereits im Wörterbuchnetz unter http://woerterbuchnetz.de/LmL nachzuschlagen, dahinter liegt, für den Nutzer nicht sichtbar, eine kodierte TEI-Datei.

Transformation in PDF- oder (X)HTML-Dokumente. Allerdings sind Auszeichnungssprachen wie TEI relativ „gesprächig“: Sie benötigen vergleichsweise viel Speicherplatz und effiziente Verarbeitungsmethoden.

Je besser Volltexte strukturiert und ausgezeichnet sind, desto hochwertiger sind die Suchindizes, die aus ihnen maschinell erzeugt werden können. Ein Index ist eine elektronische Speicherstruktur, die häufig in Form einer invertierten Liste implementiert wird. In der Liste ist für jedes Wort notiert, in welchen Wörterbucheinträgen es an welcher Position auftritt. Bei der Indizierung werden Begriffe auf ihre Grund- bzw. Stammform reduziert und Stoppwörter wie etwa Konjunktionen getilgt.

Digitalisierung von Wörterbüchern an der Akademie

Der Medienwandel hält auch in der Bayerischen Akademie der Wissenschaften Einzug: Fachkräfte werden eingestellt, dauerhafte Strukturen und Konzepte für die Digitalisierung erarbeitet, Projekte mit Kooperationspartnern in Angriff genommen. Das Lexicon musicum Latinum medii aevi, ein Wörterbuch der lateinischen musikalischen Fachsprache des Mittelalters, war weltweit eines der ersten Wörterbuchunternehmen, das von Beginn an konsequent die elektronische Datenverarbeitung für die Erfassung der Quellentexte und die Publikation nutzte. Die Auszeichnung mit TEI, die das Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften an der Universität Trier durchführte, war daher relativ leicht möglich. Die ersten Faszikel (Buchstabe A–D) stehen im Rahmen des Wörterbuchnetzes bereits online zur Verfügung unter http://woerterbuchnetz.de/LmL (Abb. 1).

Das Kompetenzzentrum hat auch die ersten im Bleisatzverfahren gedruckten Lieferungen des Mittellateinischen Wörterbuchs (s. S. 40–42) durch Abschreiben retrodigitalisiert. Außerdem wurden in Zusammenarbeit mit der Bayerischen Staatsbibliothek (BSB) die Daten des Repertoriums Geschichtsquellen des deutschen Mittelalters mit TEI ausgezeichnet und jüngst im Netz publiziert unter www.geschichtsquellen.de. Weitere Projekte mit der BSB sind in Planung.

Medien wandeln sich. Aber Medien sind lediglich Vermittler von Inhalten, die nur menschliche Experten erarbeiten können. Elektronische Werkzeuge und Medien zu schaffen, die bei der wissenschaftlichen Arbeit helfen, das ist das Ziel der Digitalisierungsgruppe in der Bayerischen Akademie der Wissenschaften.