arbido

2011/3 Elektronische Bibliothek Schweiz

Padlina Roberta,

Rüegg Monika ,

Tags: Dokumentation Analog, Elektronisch,

e-codices: Virtuelle Handschriftenbibliothek der Schweiz Informationsverarbeitung mit Metadaten

Seit 2005 digitalisiert e-codices mittelalterliche und frühneuzeitliche Handschriften aus schweizerischen Bibliotheksbeständen. Als Teilprojekt von e-lib.ch: Elektronische Bibliothek Schweiz konnte sich das Freiburger Projekt in den letzten Jahren als zentrales Handschriftenportal für die Erschliessung von Handschriften etablieren und wird vor allem auch im Ausland als modellhafte digitale Bibliothek verstanden.

In drei früheren Artikeln (*arbido* 10/ 2005, 1/2006, 3/2009) hatte e-codices schon die Möglichkeit, das Projekt vorzustellen. In diesem Artikel soll ein zentraler Aspekt des Projekts behandelt werden: die Metadaten. Metadaten werden nicht nur für die optimale Archivierung von Informationen erstellt, sondern sie ermöglichen vor allem die Programmierung von Funktionalitäten (insbesondere der Suchfunktion) und den Austausch von Informationen. Dabei beschränkt sich e-codices nicht nur darauf, bereits vorhandene Informationen zu den Handschriften möglichst unverändert wiederzugeben und in einer leistungsfähigen Datenbank durchsuchbar zu machen, sondern es werden auch neue Informationen generiert, welche die Nutzung der Daten vielfältiger und dynamischer machen. Im Folgenden soll gezeigt werden, welche Navigations- und Suchwerkzeuge e-codices anbietet und mit welchen Technologien diese umgesetzt werden.

Die Informationssuche mittels der Browsefunktion

e-codices bietet dem Benutzer mittels der Such- und der Browsefunktionen zwei sich ergänzende Möglichkeiten, Informationen zu finden. Die Funktionen unterscheiden sich in ihren verwendeten Werkzeugen und Zielen: Entweder sucht man nach einem präzisen Begriff, zum Beispiel einem bestimmten Autor mittels einer Suchmaschine (Suchfunktion), oder man «durchstöbert» verfügbare Autoren dank der Navigation durch systematische Indizes (Browsefunktion).

Manchmal weiss man generell, was man suchen möchte, weiss aber nicht, wie man es finden könnte. Die Browsefunktion hilft dem Benutzer, in der Informationsmasse Orientierungen zu finden und ermöglicht so neue Entdeckungsmöglichkeiten. Möchte der Benutzer sich generell über den Inhalt der digitalen Bibliothek informieren, so kann er - ähnlich wie jemand Bibliotheksregale durchstöbert – die Browsefunktion nutzen. Das Browsen oder eben Durchstöbern wird auf e-codices durch die Funktion «auflisten nach» ermöglicht.

Eine Dropdown-Liste gestattet es dabei dem Benutzer, eine Liste sämtlicher verfügbaren Bibliotheken anzusehen. Er hat die Wahl, eine bestimmte Sammlung auszuwählen oder alle gemeinsam zu durchstöbern. Wahlweise kann der Benutzer in einer zweiten Dropdown-Liste alle Handschriften oder diejenigen einer bestimmten Sammlung nach verschiedenen Kriterien sortieren. Zurzeit bietet e-codices folgende Sortiermöglichkeiten an: Sortieren nach *Signatur* (Standardeinstellung), nach *Autoren,* nach *Entstehungszeit,* nach *Online seit* und zusätzlich die Anzeige aller Handschriften des *letzten Updates.* So hat der Benutzer beispielsweise bei den Autoren die Möglichkeit, anhand einer alphabetischen Liste den gewünschten Verfasser eines Werkes auszuwählen. Ein weiterer Vorteil der Browseansicht sind die Kurzbeschreibungen zu jeder einzelnen Handschrift. Sie enthalten in kürzester Form die wichtigsten Angaben zum Inhalt und zur Bedeutung der Handschrift. Da diese Angaben für e-codices verfasst und konsequent in vier Sprachen angeboten werden, erhöhen diese Metadaten den Informationsgehalt und laden den Benutzer zum Durchstöbern ein.

Die Suchfunktion

Neben dem Browsen kann der Benutzer von e-codices auch ganz gezielt nach Informationen suchen. Er durchstöbert dann nicht wie beim Browsen die virtuelle Bibliothek, sondern gibt in einen Suchschlitz oder eine Suchmaske einen Begriff ein. Durch einen Klick auf den Button «Suche» kann der Benutzer so jederzeit einen konkreten Begriff in den Handschriftenbeschreibungen ermitteln. Neben der Suche im *Volltext* kann auch gezielt in einzelnen Bereichen wie *Standort/Signatur, Autor, Handschriftentitel, Werktitel, Entstehungsort, Incipit, Explicit* und *Buchschmuck* gesucht werden. Bei der Detailsuche ist besonders die Suchmöglichkeit nach Autoren erwähnenswert. Es besteht hier die Möglichkeit, den Namen des gewünschten Verfassers in das Suchfenster einzugeben. Dabei sollte es im Prinzip keine Rolle spielen, in welcher Sprache und in welcher Namensform der Autor eingegeben wird. Im Idealfall sollte der Benutzer jeweils leicht zum gesuchten Autor geführt werden: die Suche nach «Jerome», «Girolamo», nach «Jérôme» oder «Stridonius» führt so zum Bsp. zu «Hieronymus, Sophronius Eusebius (345–420)».

Die kategorisierte Informationssuche mit Facetten

Browsen und Suchen sind zwei unterschiedliche Vorgehensweisen, umgezielt Informationen zu finden. Mit sogenannten Facetten, welche in den Bibliotheksportalen immer häufiger Anwendung finden, werden die beiden Vorgehensweisen gewissermassen miteinander kombiniert. Facetten gruppieren Suchresultate in Kategorien und geben in Klammern die Anzahl gefundener Dokumente zu jeder Kategorie an. Dadurch lassen sich Listen mit vorhandenen Kategorien erstellen, welche dem Anwender ein «Browsen» in den vorhandenen Daten ermöglichen. Es lassen sich aber auch die bereits durch die Suchfunktion gefundenen Treffer weiter eingrenzen (bei e-codices nach den *Kategorien Entstehungszeit, Beschreibstoff, Ort/Bibliothek* und *Art der Beschreibung*), womit die Facetten auch die Funktion eines Suchfilters übernehmen können.

Informationen suchen und finden ist nicht immer einfach. Die Facetten sind ein gutes Beispiel dafür, wie die Technologie dem Benutzer Orientie- rungshilfen anbietet: Was gefunden werden kann, wird klar und übersicht- lich dargestellt. So kann man z.B. alle Handschriften, die Werke von Augustinus enthalten, aus dem 9. Jahrhundert stammen und nur in St. Gallen aufbewahrt sind, schnell finden.

XML – die Sprache der Struktur und der Semantik

Die Suchwerkzeuge werden ermöglicht dank einer technischen Vorarbeit, welche an den Handschriftenbeschreibungen geleistet wird. Die Arbeit mit den Daten findet auf drei Ebenen statt: der Ebene der Struktur, des Datenverhaltens und der Präsentation. Die Struktur ist die logische Ordnung, mit der die Information zerlegt und neu organisiert wird. Einen Inhalt zu strukturieren, bedeutet aber nicht nur, ihn zu zerlegen, sondern auch semantisch zu kennzeichnen, so dass die einzelnen Teile miteinander in Beziehung treten können.

Bei e-codices werden die wissenschaftlichen Handschriftenbeschreibungen nach dem Standard der Text Encoding Initiative (TEI) in der aktuellen Version P5 als XML kodiert. TEI ist ein Konsortium, welches einen Standard für die digitale Präsentation von Texten definiert und unterhält. Dessen Richtlinien stellen eine umfassende XML-Grammatik zur Verfügung. Die seit November 2007 freigegebene Version P5 enthält unter anderem auch einen Standard zur Beschreibung von Handschriften: das *Manuscript Description Modul* der TEI enthält alle nötigen Tags, um Struktur und Elemente einer wissenschaftlichen Handschriftenbeschreibung zu kennzeichnen. TEI hat sich mittlerweile nicht nur zu einem Standard der Geisteswissenschaften entwickelt, sondern konnte sich bei den meisten Handschriftendigitalisierungsprojekten durchsetzen (http://www.tei-c.org/Activities/Projects).

Bekanntlich ist XML eine Auszeichnungssprache, die der strukturellen Beschreibung des Inhalts von Dokumenten dient. Mit ihr ist es möglich, die Struktur eines Dokuments präzise zu definieren und sie mit semantischer Bedeutung anzureichern. In der Praxis heisst das, dass wir dank XML den Daten in den Handschriftenbeschreibungen neue Informationen, genauer gesagt Informationen über Informationen (auch Metainformationen oder Metadaten genannt), hinzufügen.

XML arbeitet mit Tags, die es erlauben, Teile eines Dokuments wie *Titel, Kapitel, Absatz* etc., aber auch beliebige inhaltliche Elemente wie *Personennamen, Orte, Daten, Materialien* etc. logisch zu kennzeichnen. In strukturierter Form können die Daten nicht nur konsultiert, sondern auch dazu benutzt werden, um neue Information zu generieren.

Datenbank und Programmierung: das Datenverhalten

e-codices beschränkt sich nicht darauf, den Inhalt der in gedruckter Form vorliegenden Beschreibung zu reproduzieren. Die wichtigste Arbeit liegt in der Programmierung des Verhaltens dieser Inhalte, der zweiten Ebene der Informationsverarbeitung. Die Kodierung in XML erlaubt die Automatisierung von verschiedenen Prozessen, z.B. die Generierung von Indizes. Dank der XML-Tags kann ein Computerprogramm die auf bestimmte Weise ausgezeichneten Dokumente verarbeiten: so kann z.B. ein Autorenindex automatisch generiert werden, indem das Programm aus jedem mit dem entsprechenden Tag «author» gekennzeichneten Personennamen den anzuzeigenden Text extrahiert.

XML erlaubt ausserdem eine einheitliche Archivierung der strukturierten Daten. Die Metadaten können mit Hilfe von Skripten automatisch in Datenbanktabellen umgewandelt werden, welche wiederum die Basis für die Implementierung der Browsefunktionen bilden. Die Ebene des Verhaltens setzt die Ebene der Struktur voraus, denn ohne diese wäre jede Art von Weiterverarbeitung gar nicht erst möglich.

Zur Verarbeitung der Daten und Metadaten in einer relationalen Datenbank wird die Sprache SQL *(Structured Query Language)* eingesetzt. Die Daten sind in Tabellen abgelegt und über sogenannte Relationen (Beziehungen) miteinander verknüpft. Die Granularität der Daten ist dabei von grosser Bedeutung: je feiner die Daten aufgeteilt und beschrieben werden, desto grösser wird ihr Benutzungspotential.

Die von e-codices verwendete Datenbanklösung heisst MySQL und kommuniziert mit der Programmiersprache PHP. Eine vertiefte und exakte wissenschaftliche Recherche nach selektiven Kriterien wird durch die Indexierung der Metadaten ermöglicht. Als interne Suchmaschine für die Suche in diesem Index verwendet e-codices *Solr,* einen Suchserver, der mit der Suchsoftware Lucene arbeitet. Alle von e-codices eingesetzten Technologien sind Open Source Software.

Eine Programmiersprache definiert nicht die Struktur oder das Aussehen von Informationen, sondern vor allem die Operationen, die mit den Informationen ausgeführt werden. Dank ihr lassen sich die Daten, welche in unterschiedlichen Strukturen enthalten sind, nach Massgabe der Bedürfnisse der Suche identifizieren, extrahieren und neu organisieren. Die Programmierarbeit wird hierbei für den Benutzer nicht ersichtlich, weil er ein bereits fertig verarbeitetes Dokument zur Ansicht erhält.

So kann beispielsweise in der Autorendatenbank der Name eines Verfassers mit einer spezifischen Identifikationsnummer (Personennamendatei Nummer oder «PND»-Nummer) verbunden werden. Damit ist gewährleistet, dass die Suche nach einem Autor mit verschiedenen Übersetzungsvarianten immer zu einem eindeutigen Ergebnis führt. Um das oben genannte Beispiel aufzugreifen: Der Benutzer gibt «Jérôme» in die Suchmaske ein, dieser Name wird intern mit der PND-Nummer 118550853 verbunden, und der Benutzer wird direkt zu den Werken von Hieronymus geleitet.

Die Präsentation der Metadaten

Die Metadaten sind auch das Schlüsselelement für die Präsentation der Beschreibungen im Internet. Mit Hilfe der Transformationssprache XSL werden Formatierungsregeln in sogenannten Stylesheets definiert, welche auf die Metadaten angewandt werden, so dass diese in ein beliebiges Format, zum Beispiel eine HTML-Webseite, umgewandelt werden können. XSL ermöglicht damit die dritte Ebene der Informationsverarbeitung, nämlich deren Präsentation. Wie das Verhalten hängt auch die Präsentation der Daten von ihrer Struktur ab. Diese wird umso präziser und ausgefeilter, je detaillierter und tiefer die Struktur ist.

Dank der Metadaten können Informationen von grossem Umfang auch *gemeinsam nutzbar* und *wiederverwendbar* gemacht werden. Dank der Kodierung mit XML nach dem TEI-P5-Standard können Metadaten auch für andere Plattformen zugänglich gemacht werden. Gerade weil die XML-Daten zwar strukturiert sind, aber keine stilistischen Eigenheiten aufweisen, können sie weitergegeben oder wiederbenutzt werden. Auch hier gilt ganz allgemein: Je tiefer und präziser ein Dokument «getaggt» ist, desto mehr Möglichkeiten bestehen für die Wiederverwendung und den Austausch der Inhalte und entsprechend höher ist das kommunikative Potential eines Dokuments. Die Stärke von XML liegt gerade in der Möglichkeit, die Daten für andere Zwecke und in verschiedenen Bereichen wiederzuverwenden, wobei dadurch auch neue Informationen generiert werden können. Dank der Verknüpfung von Informationen können neue Zugangskanäle geschaffen werden, welche die gemeinsame Nutzung von Kenntnissen, Informationen und Dokumenten unterstützt. Diese Einfachheit der gemeinsamen Nutzung von Informationen ermöglicht eine bessere Vernetzung und erhöht potentiell die Qualität der wissenschaftlichen Forschung.

Zukunftsaussichten

Die Webtechnologien unterliegen einem stetigen Wandel. Neue Ansätze entstehen wie Linked Data und neue Akronyme werden geläufig wie RDF. Mit dem sogenannten «Semantic Web» kündigt sich nach dem viel diskutierten Web 2.0 bereits das Web 3.0 an. Dabei geht der Trend in Richtung einer universellen semantischen Klassifizierung, so dass Beziehungen zwischen Objekten, Konzepten und Inhalten im Netz hergestellt werden können. Durch das Internet wurden die Archivierung, die Verteilung und die Nutzung von Daten von Grund auf verändert. Diese Entwicklung wird auch in Zukunft weiter voranschreiten. Informationsangebote wie e-codices müssen versuchen, mit dieser Entwicklung Schritt zu halten und die Website den wachsenden Bedürfnissen und Anforderungen der Benutzer anzupassen. Es ist unsere Aufgabe, durch die neuen Möglichkeiten, welche die Technologie zur Verfügung stellt, einen optimalen Zugang zur Information zu leisten.

Konkret betrachtet sind die Möglichkeiten der Facettierung, also der Zuordnung von Such-, Browse- und Filterkategorien, noch nicht ausgeschöpft und sollen in naher Zukunft noch erweitert werden. Beispielsweise könnte der Benutzer dann die Handschriften nach verschiedenen Disziplinen wie Musikwissenschaft, Philosophie, Liturgiewissenschaft oder nach Herkunft bzw. Provenienz durchsuchen. Eine übersichtliche und leistungsfähige Bildsuche zu schaffen, ist nach wie vor eine der grössten und schwierigsten Herausforderungen. Aber auch die Verknüpfung mit weiteren Websites und Portalen wird angestrebt. Es ist geplant, dass schon bald mittelalterliche Skriptorien, deren Handschriften heute in alle Welt zerstreut sind, wieder virtuell zusammengeführt werden können. Dasselbe gilt für Handschriften, die in Einzelteile zerlegt wurden und nun in verschiedenen Institutionen aufbewahrt werden.

e-codices ist in internationale Bemühungen eingebunden, mittels Linked Data einen einheitlichen und interoperablen Standard für die Präsentation und Nutzung von virtuellen (Handschriften-)Bibliotheken zu schaffen. Durch die Einbindung von Annotations- und Transkriptionswerkzeugen, die zurzeit an mehreren Universitäten im In- und Ausland entwickelt werden, wird e-codices als nationales Handschriftenportal und Teil der Elektronischen Bibliothek der Schweiz immer mehr zu einer Plattform werden, die nicht nur Primärquellen zur Verfügung stellt, sondern potentiell alle zu einer Handschrift gehörigen Forschungsergebnisse versammelt und so selbst zu einem (virtuellen) Ort der internationalen Handschriftenforschung wird.

Roberta Padlina

Projekt e-codices, Universita?t Fribourg

Monika Rüegg

Projekt e-codices, Universita?t Fribourg

Abstract

Français
Depuis 2005, e-codices procède à la numérisation de manuscrits médiévaux et du début de l'époque moderne, conservés dans les bibliothèques suisses.

Inscrit dans le cadre de e-lib.ch, le projet fribourgeois a pu s'imposer ces dernières années comme un portail central pour la mise en valeur des manuscrits et est vu essentiellement, également à l'étranger, comme un modèle de bibliothèque numérique. Dans trois articles précédents (Arbido 10/2005, 1/2006, 3/2009), e-codices a eu l'occasion d'être décrit. La présente contribution s'intéresse à un aspect central du projet: les métadonnées. Les métadonnées sont créées non seulement pour un archivage optimal d'informations, mais permettent également la programmation de fonctions (en particulier de recherche) et autorisent l'échange d'informations.

Pour cette raison, e-codices ne se limite pas à restituer l'information existante sur les manuscrits dont l'accès serait assuré par une puissante base de données. Le projet sert également à générer de nouvelles informations, qui permettent une approche dynamique des données. L'article expose les modes de navigation et les outils de recherche offerts par e-codices, et précise les technologies mises en œuvre.