"Digitalisierung fremdschriftlicher Texte bei Anwendung normierter Transkriptionssysteme"

Hanstein, Thoralf ; Kupferschmidt, Jens; Scharsky, Alfred

Problematik

Die informationstechnische Erfassung und Aufarbeitung alter Handschriften, die in einer Sprache mit nichtlateinischem Alphabet geschrieben worden sind, stellt noch immer eine große Herausforderung an alle Beteiligten dar. Die kodikologischen Beschreibungen müssen in verschiedenen Schriften und Sprachen als Metadaten in einer Datenbank erfassbar sein.
Schnittstellen für fremdschriftliche Katalogisate und den damit verbundenen Transkriptionen zwischen Verbundsystemen und Lokalsystemen stellen noch immer eine große Herausforderung dar.

Es gibt einige Sprachen, die mit dem arabischen Alphabet verschriftlicht werden[1], und allen gemeinsam sind Besonderheiten wie z.B. die Nichtwiedergabe der kurzen Vokale in der Schriftform, die verschiedenen Formen eines Buchstabens - abhängig von seiner Stellung im Wort (am Anfang eines Wortes, in der Mitte, am Ende und isolierte Stellung) - und auch oft rein ästhetisch bedingte Ligaturen. Unter anderem diese Eigenheiten der arabischen Schrift erschweren deren Verarbeitung in der Informatik. Aussagekräftige Beispiele für die auftretenden Probleme sind die immer noch unbefriedigenden Ergebnisse in der Texterkennung (OCR) des Arabischen[2]. Die datenbankgestützte Beschreibung islamischer Handschriften und ihre Darstellung im Internet wurde von der DFG als übergreifende Aufgabe anerkannt und gefördert.

Im „Pilotprojekt zur datenbankgestützten Erschließung und digitalen Bereitstellung der neu erworbenen arabischen, persischen und türkischen Handschriften der Universitätsbibliothek Leipzig“ konnte unter ausschließlicher Verwendung freier Software eine Anwendung entwickelt werden, die den vielfältigen Anforderungen der Kodikologie gerecht wurde.[3]

Bereits in der Antragsphase des Projekts wurde der internationale Stand der datenbankbasierten Katalogisierung und gleichzeitigen online-Präsentation orientalischer Handschriften evaluiert.[4] Im Ergebnis musste jedoch eingeschätzt werden, dass – wahrscheinlich oft bedingt durch die fremdschriftliche Komponente – kaum Sammlungen orientalischer Handschriften der renommierten Institutionen im Internet vertreten waren. Ausnahme war die Haddād-Handschriften-Sammlung der Wellcome Library in London[5]. Dort wurden zwar keine Bilder der Manuskripte ins Netz eingestellt, aber der datenbankbasierte Katalog beinhaltete Einträge in Arabisch und auch in Transkription mit Sonderzeichen. Leider gab und gibt es auf der Seite Darstellungsprobleme mit dem Font, sodass z.B. einige Sonderzeichen nur als Fragezeichen erscheinen.

Auch innerhalb der deutschen Arabistik und Islamwissenschaft wurden vereinzelt Anstrengungen auf diesem Gebiet unternommen, aber im Ergebnis standen immer nur inkompatible und isolierte

---

1 Als Auswahl u.a.: http://de.wikipedia.org/wiki/Liste_arabisch-basierter_Alphabete

2 u.a. in: Bazzi, I., Schwartz, R., u. Makhoul, J., An Omnifont Open-Vocabulary OCR System for English and Arabic. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, Nr. 6, Juni 1999. ;
Lorigo, L. M. u. Govindaraju, V., Offline Arabic Handwriting Recognition: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 28, Nr. 5, Mai 2006.

3 http://www.islamic-manuscripts.net, Nachfolgeprojekte: www.manuscripts-aceh.org, www.refaiya.uni-leipzig.de

4 Omar El Bannay, Noureddine Rais, Rachid Benslimane, Noureddine El Makhfi, "Application for visualization of arab manuscripts," aiccsa, pp.779-782, 2009 IEEE/ACS International Conference on Computer Systems and Applications, 2009

5 http://library.wellcome.ac.uk/node273.html

Insellösungen. Eine gängige pragmatische Herangehensweise bestand im Scannen der gedruckten Kataloge, wobei aber infolge der verschiedenen Sprachen und Alphabete nicht mit Hilfe von OCR durchsuchbare Varianten entstanden, sondern rein grafische Darstellungen der kodikologischen Beschreibungsdaten.

Bereits in der Vorbereitungsphase des Projekts wurde sehr schnell klar, dass eine einfache relationale Datenbank als Basis nicht in Frage kam, da die Abbildung einer Mehrsprachigkeit hier nur mit sehr komplexen Mitteln realisiert werden kann.

Mehrsprachigkeit und Mehrschriftlichkeit unter besonderer Berücksichtigung des Arabischen

Die besondere Herausforderung der Arabischen Sprache an die Softwarepräsentation ist zweifellos die rechts-links-Schreibrichtung (rtl). Auch kulturelle Aspekte wie z.B. besondere Lesegewohnheiten und Gestaltungskonstanten müssen beim Design einer Datenbankoberfläche beachtet werden.

Durch die konsequente Nutzung von XML[6] mit der Transformation durch XSLT[7] konnte das Datenmodell mit mehrsprachigen Werten somit je nach ausgewählter Anwendungssprache dargestellt werden. Das gilt nicht nur für die eigentlichen Daten, sondern auch für die statischen Komponenten wie Navigation und beschreibende Webseiten. Ergänzt wird diese Technologie durch den Einsatz der I18N-Technik. Mit diesem Verfahren lassen sich Label, Hilfetexte usw. in den erforderlichen Sprachen gestalten.
Die Eingabemaske für die kodikologischen Beschreibung können in Deutsch, Englisch, Indonesisch und Arabisch aufgerufen werden. Die Präsentation auf der Webseite erfolgt automatisch in der vom Nutzer gewählten Navigationssprache. Auch der spezielle Bildbetrachter IView ermöglicht durch seine Flexibilität die Anpassung an die Besonderheiten der Ausrichtung: die Scans werden in der Thumbnailübersicht von rechts nach links in der den Arabern gewohnten Lesereihenfolge dargestellt.

Wissenschaftliche Transkription

Eine weitere Herausforderung bestand in der Implementierung der verschiedenen Transkriptionsysteme. Für orientalische Sprachen sind als gebräuchlichste das in Deutschland verwendete System der Deutschen Morgenländischen Gesellschaft (DMG)[8] und das anglo-amerikanische System der Library of Congress (LC)[9] zu nennen.
Gerade die für die wissenschaftliche Transkription benötigten Sonderzeichen bereiteten den

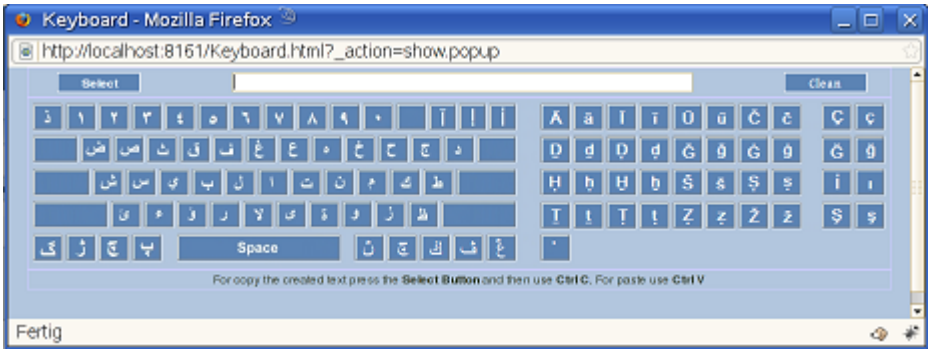
6 Extensible Markup Language, siehe http://www.w3.org/XML/
7 Extensible Stylesheet Language Transformation, siehe http://www.w3.org/TR/xslt20/
8 Carl Brockelmann: Die Transliteration der arabischen Schrift in ihrer Anwendung auf die Hauptliteratursprachen der islamischen Welt. Denkschrift, dem 19. internationalen Orientalistenkongreß in Rom vorgelegt von der Transkriptionskommission der DMG (Deutsche Morgenländische Gesellschaft) /von Brockelmann, Carl; Fischer, August; Heffening, W.; Taeschner, Franz mit Beiträgen von Ph. S. van Ronkel und Otto Spies. DMG in Kommission bei F.A. Brockhaus: Leipzig 1935.
9 http://www.loc.gov/catdir/cpso/roman.html

Arabisten und Islamwissenschaftlern bei der Textverarbeitung und im Internet bisher einige Probleme. Lange Zeit wurden Kompromisse bei der Darstellung der Sonderzeichen im Internet in Kauf genommen. Dies geht zumeist auf Kosten des wissenschaftlichen Anspruchs bzw. auch der Nutzerfreundlichkeit.

Durch die immer konsequentere Anwendung des Unicode-Standards in den letzten Jahren hat sich jedoch Vieles vereinfacht und vereinheitlicht. Proprietäre Browser-Lösungen wie z.B. Microsofts Weft *(*Web Embedding Fonts Tool*)* sind damit zumindest in Bezug auf die Transkription orientalischer Texte nicht mehr relevant. Immer mehr Standardfonts beinhalten das komplette Set der benötigten Sonderzeichen, sodass jetzt im Prinzip nur noch die Frage der nachträglichen Umformatierung der alten Texte in nichtunicodefähigen Umschriftsfonts auf effektive Weise gelöst werden muss.
Im Pilotprojekt wurde die Auswahl des Transkriptionssystem an der aktuellen Navigationssprache festgemacht: Geht man über die deutsche Webseite zu den Handschriften, wird automatisch die Umschrift der DMG angeboten; bei der englischen Navigation erscheint die Umschrift der LC.

Eingabehilfe für Sonderzeichen

Auch in Bezug auf Nutzerfreundlichkeit bei der Eingabe der Sonderzeichen oder auch der arabischen/persischen Sprache an sich konnte im Pilotprojekt eine Lösung aufgezeigt werden. Es wurde ein systemunabhängiges onscreen-Keyboard auf Basis von HTML und Java Script entwickelt, mit dem in der aktuellen Version neben den erforderlichen Transkriptionssonderzeichen auch Buchstaben und Wörter in Arabisch, Persisch, (Osmanisch-)Türkisch und Jawi[10] eingegeben werden können. Erweiterungen für Javanisch und weitere Alphabete sind in Vorbereitung.



Flexible Suchmöglichkeiten

Letzte - aber keineswegs in der Bedeutung nachrangige - Herausforderung war die Umsetzung einer optimalen Suchumgebung für den Nutzer der Datenbank. Auch wenn die korrekte Eingabe der Daten in der Originalsprache und in Transkriptionssystemen wichtige Grundlagen für die wissenschaftliche Arbeit ist, so ist doch die Nutzerfreundlichkeit vor allem in Bezug auf das Auffinden gesuchter Informationen das primäre Ziel. In der Projektlösung wurden deshalb mehrgleisige, flexible Suchmöglichkeiten eingerichtet. Der Nutzer kann somit zwischen der Volltextsuche in den Metadaten, der Filterung der Daten über definierte Klassifikationen und dem direktem Indexbrowsing wählen. Des Weiteren können über spezielle Suchmasken Kombinationen der genannten Suchstrategien zusammengestellt werden. Hintergrund bildet ein Daten-Indexer aus dem Apache Lucene Project, welcher in Kombination mit der Nutzung eines Textnormalisierers eine effiziente Suche gestattet.
Bei der Freitextsuche kann der Nutzer somit das gesuchte Wort z.B. direkt in Arabisch eingeben oder zwischen den beiden Transkriptionssystemen wählen. Zusätzlich wurde aber auch eine

10 Malaiisch mit arabischen Buchstaben geschrieben.

sogenannte „Normalisierung“ der Sonderzeichen eingerichtet. Dabei kam die ICU[11]-Implementierung der Firma IBM zum Einsatz. Die erforderliche Bibliothek unterliegt einer freien Lizenz für nichtkommerzielle Anwendungen. Durch die Normalisierung können die Sonderzeichen bei der Eingabe des Suchworts auch ignoriert werden: die Suche nach البخاري = Buḫārī = Buhari bringt dasselbe Ergebnis.

Da viele Namens- und Ortsangaben in unterschiedlichen Schreibweisen angegeben werden, die nicht durch Normalisierung abgedeckt werden können, wird für die Weiterentwicklung des Projektes die Integration von Thesauri für die Suche angestrebt.

Gestaltung der Administrationseben

Neben der Darstellung der Inhalte im Internet war es den Entwicklern auch besonders wichtig, alle zur Betreuung erforderlichen Arbeiten möglichst online ausführen zu können. Hier wurden zum einen die bereits mit dem MyCoRe-Kern bereitgestellten Funktionen verwendet, die meist einer Administration des Gesamtprojektes dienen (z. B. User, Commandline Tool usw.). Zum anderen werden Editormasken für die Dateneingabe der Metadaten und Java-Applets zum Einstellen der Bilder angeboten. Damit können die Kodikologen bzw. Hilfskräfte standortunabhängig an der Datenbank arbeiten. Erforderlich hierfür ist lediglich ein Web-Browser, bevorzugt ist Firefox empfohlen.
Die Anwendung gestattet dabei ein feingranulares Rechtesystem für den Zugang zu Bearbeitungsfunktionen. Dabei können mehrere Klienten unabhängig und ohne sich gegenseitig beeinflussen zu können, gleichzeitig ihre Daten verwalten. Auch für die Recherche und Navigation kann zwischen der für einen einzelnen Klienten und der für das Gesamtsystem gewählt werden.

Architektur der Anwendung

Inzwischen entstanden aus dem ursprünglichen Prototyp eine ganze Reihe von Anwendungen desselben Typs. Lediglich kleinere Teile wie das farbliche Layout, Texte zu den Sammlungen oder die Anzahl der angebotenen Sprachen sind dabei different. Der Großteil der Anwendung ist dabei gleich. Aus dem ursprünglichen Prototypen IslamHS wurde eine Basisanwendung mit Namen MyIHS[12] extrahiert, welche auch eigenständig als Demo-Version alle Features der Applikationsgruppe aufzeigen kann. Als neue Projekte wurden daraus bisher entwickelt:

- IslamHS – als Weiterentwicklung des ursprünglichen Prototyps
- Refaiya – ein DFG-gefördertes Projekt zur Erschließung der Privatbibliothek Refaiya an der UBL
- AcehMS – ein Projekt zur Katalogisierung und Digitalisierung von Handschriften in der Region Aceh / Indonesien
- JavaMS - ein Projekt zur Katalogisierung und Digitalisierung von Handschriften auf Java / Indonesien

Für weitere Projekte sind die on-the-fly-PDF-Generierung von Mischtexten mit Arabisch zu Dokumentationszwecken und die Erweiterung um einen Datentyp zur Beschreibung von Eignern der Handschriften vorgesehen.

Prinzipiell lassen sich die Erfahrungen mit Texten aus dem arabischen Bereich auch auf andere nichtlateinische Schriften und Transkriptionssysteme übertragen.

---

11 International Components for Unicode

12 Die Gesamte Software beginnend ab MyCoRe unterliegt der GNU Public License 2 und ist für nichtkommerzielle Zwecke als Open Source Projekt kostenfrei nutzbar. Aus Gründen der Qualitätssicherung werden downloadbare Distributionen nicht im Netz direkt angeboten. Interessenten können aber Anfragen jederzeit per Mail an collection@rz.uni-leipzig.de senden. Daneben sind die Codes über ein Subversion-System (http://svnextern.dl.uni-leipzig.de/) auch direkt verfügbar.