

ALO oder die virtuelle Bibliothek der österreichischen Literatur

Ein Arbeitsbericht

Alexander Egger, Günter Mühlberger

1. Einleitung

1.1 Ausgangslage

Weltweit beschäftigt sich eine Vielzahl von Projekten und Programmen mit der Retrodigitalisierung von Büchern und Zeitschriften. Diese Anstrengungen sind inzwischen für viele Bibliotheken und Länder zu wichtigen „Aushängeschildern“ ihrer kulturellen Identität in einer globalen Informationsgesellschaft geworden. Hinzu kommt, dass auch Medien in einem Konkurrenzverhältnis zueinander stehen und bereits heute absehbar ist, dass künftig nur noch jene

Information in der Öffentlichkeit wahrgenommen wird, die in elektronischer Form vorliegt. Die digitale Konversion von Büchern und Zeitschriften ist somit auch eine aus kultur- und wissenspolitischer Sicht eminent wichtige Aufgabe. Ein Blick ins Internet zeigt jedoch, dass trotz der vielen Millionen Buchseiten, die bisher digital konvertiert wurden und im Internet der Öffentlichkeit zur Verfügung stehen, nur ein verschwindend geringer Anteil die deutschsprachige oder gar österreichische Literatur umfasst. Diesem kulturellen Defizit soll nun mit der Initiative für ALO (*austrian literature online - österreichische literatur online*) entgegengetreten werden. Das plakative Ziel von ALO: Die 1000 wichtigsten Bücher der österreichischen Literatur vom Ende des 18. Jahrhunderts bis 1930 (Urheberrecht) sollen digitalisiert und der Öffentlichkeit zugänglich gemacht werden. Die ausführliche Darstellung der Ergebnisse der Machbarkeitsstudie sowie der ausgewählten Beispiele sind im Internet abrufbar: <http://alo.aib.uni-linz.ac.at/>.

1.2 Machbarkeitsstudie und Prototyp

Im Frühjahr 1999 wurde daher von einer Arbeitsgemeinschaft österreichischer Bibliotheken und Universitäten beim Bundesministerium für Wissenschaft, Bildung und Kultur eine Machbarkeitsstudie beantragt, die sich zum Ziel gesetzt hat, die Voraussetzungen für eine Verwirklichung einer virtuellen Bibliothek der österreichischen Literatur zu prüfen und entsprechende Empfehlungen zu erarbeiten. Zusätzlich zur Studie wurde auch eine einfache technische Installation erarbeitet, die die wesentlichen Funktionen einer virtuellen Bibliothek beinhaltet und als Referenz für ein zukünftiges Modell dienen kann. Dieser Prototyp einer virtuellen Bibliothek der österreichischen Literatur umfasst:

- Digitale Konvertierung einiger für die virtuelle Bibliothek der österreichischen Literatur repräsentativer Bücher und Zeitschriften
- Speicherung der elektronischen Faksimiles, des automatisch erkannten Volltextes und der für den Zugriff und die künftige Datenintegrität notwendigen Metadaten (MOAll-Modell)
- Navigation in den digitalen Faksimiles, Suche im Volltext und den Indexdaten
- Herstellung verschiedener elektronischer Ausgabeformate für unterschiedliche Einsatzbereiche, z.B. PDF-Dateien für den Ausdruck oder Audio-Dateien für Blinde oder sehbehinderte Personen
- Integrierter Reprint eines ausgewählten Buches (Kafka: Hungerkünstler 1924) über das Book-on-Demand-Service der Firma *Libri*

- Zugriff auf die elektronischen Bücher mittels des österreichischen Verbundkatalogs ALEPH <<http://BVZR.bibvb.ac.at:4505/ALEPH>>.

1.3 Vom Prototyp zur etablierten Einrichtung

Mit dem Abschluss des Vorprojekts soll der Startschuss gefallen sein, für die Verwirklichung der *ALO-Bibliothek* auf breiter Basis. In nächster Zeit werden die Betreiber der „österreichischen literatur online“ daher Schritte setzen, um möglichst viele private und öffentliche Geldgeber zu überzeugen, einen Beitrag für den Ausbau und die Vergrößerung der virtuellen Bibliothek der österreichischen Literatur zu leisten. So könnten einzelne Landeskulturabteilungen oder Kulturämter größerer Städte die digitale Konvertierung der für ihr Land oder ihre Gemeinde wichtigsten Bücher finanzieren bzw. über das ALO-System abwickeln. Für private Sponsoren ist an die Einrichtung von „Bücherpaketen“, also thematisch klar umrissene Bücherkonvolute, die mit dem Aufgabengebiet des jeweiligen Sponsors in einem engen Zusammenhang stehen, gedacht. Um ein Beispiel für ein derartiges Bücherpaket zu nennen: „Schriften zur Pressefreiheit in Österreich - die 100 wichtigsten Bücher von 1780 bis 1930“. Als Sponsoren wären Zeitungen, Medienbetriebe oder Journalistenverbände denkbar.

1.4 Digitale Konversion - neue Kontexte

Neben den engeren Zielen des Projekts einer virtuellen Bibliothek der österreichischen Literatur ist auch an verstärkte Bemühungen für eine Integration der digitalen Konversion in bestehende Arbeitsabläufe gedacht. Im Bereich der Fernleihe und der Unterstützung von Forschungsprojekten könnte die digitale Konversion von Büchern und Zeitschriften bereits in nächster Zeit einige Bedeutung gewinnen.

Erstens zum Thema *Fernleihe*: Das Verschicken alter Bücher im Rahmen der Fernleihe ist ein zeitraubendes, teures und für den Zustand des Buches wenig schonendes Verfahren der Literaturversorgung. Viele Bibliotheken lehnen daher die Fernleihe älterer Bücher überhaupt ab, so dass dem Wissenschaftler oftmals nichts anderes übrig bleibt, als vor Ort zu reisen, um ältere Literatur einsehen zu können. Hier wäre eine Integration der elektronischen Konversion in die tägliche Arbeit der Bibliotheken besonders wünschenswert. Die Bestellung eines Buches könnte zu seiner automatischen Digitalisierung benutzt werden und so der Bestand an digitalisierter Literatur (im weiten Sinn) ständig erweitert werden - noch dazu mit dem Vorteil der impliziten Beachtung der Bedürfnisse der Benutzer. Da bereits heute die Gebühren einzelner Bibliotheken für die Fernleihe durchaus beachtlich sind, würden sich die zu erwartenden

den Mehrkosten für den Endbenutzer in Grenzen halten. Hingegen wäre aber das Buch der Öffentlichkeit für alle Zukunft leicht zugänglich gemacht und außerdem ein Beitrag zu seiner physikalischen Erhaltung geleistet, da nur noch in Ausnahmefällen das Buch dem Benutzer direkt vorgelegt werden müsste.

Zweitens zum Thema *Forschungsprojekte*: Eine besonders interessante Option scheint uns die gezielte Anwendung der digitalen Konversion im Rahmen bestehender und künftiger Forschungsprojekte zu sein. Die für jede historische Arbeit notwendige Literatur könnte gleich zu Beginn digital konvertiert werden und so als elektronische Forschungsbibliothek zugänglich gemacht werden. Denkbar wäre, um ein ganz konkretes Beispiel zu nennen, dass für ein Forschungsprojekt wie „Der historische Roman von 1780 bis 1945“ <<http://germanistik.uibk.ac.at/hr/>>, das in den Jahren 1991 bis 1996 vom Fonds zur Förderung wissenschaftlicher Forschung finanziert wurde, die wichtigsten 150 oder 200 Bücher möglichst zu Beginn elektronisch konvertiert werden. Damit wäre jedem Außenstehenden die Möglichkeit geboten, die Forschungsergebnisse direkt mit dem zugrunde liegenden Forschungsmaterial zu messen - ein Unterfangen, das ansonsten aufgrund der meist schweren Zugänglichkeit der Bücher unverhältnismäßig großen finanziellen und zeitlichen Aufwand erfordern würde. Angesichts der Gesamtkosten eines derartigen Forschungsprojekts wäre die digitale Konversion eine zu vernachlässigende Größe, die aber angesichts der Qualitätssteigerung für die Forschung einen unverhältnismäßig großen Nutzen erzielen würde.

2. Anforderungen an das ALO-System

Mit dem ALO-System werden elektronische Bücher erstellt, in einer digitalen Bibliothek (*ALO Library*) verwaltet und interessierten Benutzern über das Internet zur Verfügung gestellt. Die digitale Speicherung der Bücher muss ein höchstes Maß an Dauerhaftigkeit der Daten sichern. Das ALO-System verwendet daher zur Speicherung der Datei nur standardisierte Dateiformate.

Die Anforderungen an das System richten sich nach den in [MPW99] definierten Richtlinien zur Erstellung von elektronischen Texten.

2.1 Aufbau eines elektronischen Buches

Im ALO-System besteht ein elektronisches Buch aus den digitalen Faksimiles der originalen Buchseiten, dem maschinenlesbaren Text dieser Seiten sowie Metadaten über die originale und die elektronische Version des Buches. Die digitalen Faksimiles erlauben eine genaue Rekonstruktion des Originals, der maschinenlesbare Text erlaubt den Einsatz von Volltextsuch- und Textanaly-

seprogrammen und macht den Text sehbehinderten und blinden Menschen zugänglich.

Digitale Faksimiles für die Archivierung sollten mit einer Auflösung von mindestens 400 dpi und einer Farbtiefe von 24 Bit erstellt werden (siehe [MPW99]). Faksimiles dieser Art wirken fotorealistisch, benötigen aber sehr viel Speicherplatz. Sie sind daher für Darstellungen in HTML-Browsern nur schlecht geeignet. Es muss also aus den hochauflösenden Faksimiles ein zweiter Satz für die Darstellung im WWW erzeugt werden. Diese Faksimiles sollten eine Auflösung von 100 dpi und eine Farbtiefe von 8 Bit (Graustufen) haben. Damit haben sie eine für die Übertragung über das Internet akzeptable Größe. Beim Erstellen der Faksimiles muss berücksichtigt werden, dass beim ALO-System die Faksimiles mit einem Texterkennungsprogramm weiterverarbeitet werden. Texterkennungsprogramme liefern oft nur für bestimmte Auflösungen und Farbtiefen optimale Ergebnisse. Die Anforderungen an Auflösung und Farbtiefe der Faksimiles sollten daher dem verwendeten Texterkennungsprogramm angepasst sein.

Der Buchtext kann durch Abtippen oder durch ein Texterkennungsprogramm erzeugt werden. Bei der Verwendung von Texterkennungsprogrammen ist zu beachten, dass es sich bei den Büchern um alte Schriften handeln kann, die in Fraktur oder anderen alten Schriftarten gedruckt sind. Das Texterkennungsprogramm muss in der Lage sein, auch diese Schriftarten zu erkennen. Der Text sollte mit Markup versehen werden, der den von der *Text Encoding Initiative (TEI)*¹ erstellten Richtlinien folgt.

2.2 Funktionen der digitalen Bibliothek (ALO-Library)

Elektronische Bücher werden in der ALO-Library abgelegt und verwaltet. In der digitalen Bibliothek werden außerdem die Metadaten über die Bücher gespeichert und eine Liste aller verfügbarer Bücher verwaltet.

Der Zugang zur ALO-Library für Systemverwalter erfolgt über ein einfaches Management-System. Es enthält Programme, die die Erstellung von elektronischen Büchern aus den Faksimiles und dem Buchtext eines Buches unterstützen. Über dieses System werden elektronische Bücher in die ALO-Library eingefügt oder aus der ALO-Library gelöscht.

Der Benutzerzugang zur ALO-Library erfolgt über das WWW. Der Benutzer kann über HTML-Seiten nach Büchern suchen, deren Text und Faksimile be-

1 TEI <<http://www.uic.edu/orgs/tei/>> entwickelt Dateiformate für elektronische Texte. Die Formate TEI und TEIxlite (XML Version) gelten als de facto Standard.

trachten und sich die Buchdateien in diversen Formaten (PDF,² XML,³ ASCII-Text etc.) herunterladen. Benutzer, die gedruckte Bücher vorziehen, können einzelne Bücher direkt bei einem BookOnDemand-Service bestellen. Die Suchmöglichkeiten umfassen Metadaten- und Volltextsuche. Mit der Metadaten- und Volltextsuche kann nach Daten wie Autor, Buchtitel oder Schlüsselwörter gesucht werden. Bei der Volltextsuche wird der Text aller Bücher nach bestimmten Wörtern durchsucht. Ein Ziel des ALO-Projekts ist, auch behinderten Menschen den Zugang zu Literatur zu ermöglichen. Die HTML-Seiten sind daher so gestaltet, dass sie auch für blinde und sehbehinderte Menschen nutzbar sind.

3. Realisierung des ALO-Systems

3.1 Übersicht

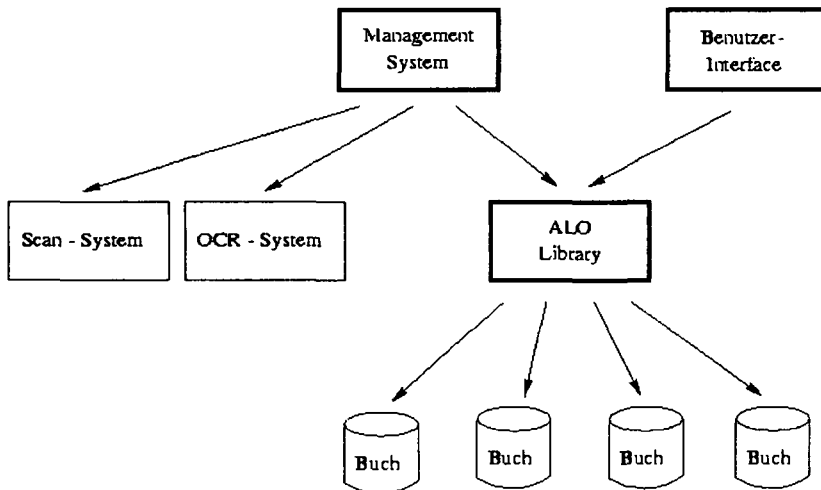


Abbildung 1: Übersicht über die Teilsysteme des ALO-Systems

Texterkennung- (OCR) und Scan-System sind externe Programme, die nicht in das ALO-System integriert sind. Der Output dieser Programme wird vom Management-System benutzt, um ein elektronisches Buch zu erstellen. Das

-
- 2 Portable Document Format. Ein von der Firma Adobe entwickeltes Dateiformat für die Ausgabe auf Druckern und Bildschirmen
 - 3 Extensible Markup Language. Markup Sprache für die Strukturierung von digitalen Dokumenten

fertige Buch wird danach in der ALO-Library gespeichert. Benutzer können dann über das Benutzerinterface auf das Buch zugreifen.

3.2 Scan- und OCR-System

Für das ALO-System steht an der Universitätsbibliothek Graz ein Minolta-Buchscanner zur Verfügung, mit dem die Faksimiles erzeugt werden. Die Bücher werden mit einer Auflösung von 600 dpi in Schwarz/Weiß gescannt und als Graphikdatei im TIFF-Format auf CD-ROM archiviert. Das Scannen der Bücher in Graustufen oder Truecolor ist mit dem Buchscanner nicht möglich. Für die Verwendung in der ALO-Library werden die Faksimiles in Dateien im GIF-Format konvertiert und auf der Festplatte des ALO-Servers gespeichert.

Als OCR-System wurde aufgrund der besten Erkennungsgenauigkeit das Programm *Finereader* gewählt. Gängige OCR-Systeme erreichen bei Vorlagen in lateinischer Schrift Erkennungsgenauigkeiten bis zu 99,8% (siehe [Ebe00]). Ältere Schriftarten wie zum Beispiel Fraktur, die im deutschsprachigen Raum in rund 80% aller Bücher vor 1942 verwendet wurde, können diese Systeme standardmäßig nicht erkennen. Allerdings können viele OCR-Programme auf neue Schriftarten trainiert werden. Für den Einsatz im ALO-System wurden verschiedene OCR-Programme auf ihre Trainierbarkeit für Fraktur getestet. Die Ergebnisse werden in den *Tabellen 1 und 2* zusammengefasst.

Readiris	Recognita	Textbridge	FineReader
120 min	125 min	55 min	75 min

Tabelle 1: Trainingsdauer

Readiris	Recognita	Textbridge	FineReader
56,36%	74,65%	93,46%	94,75%

Tabelle 2: Erkennungsgenauigkeit bei Frakturschrift

3.3 ALO-Library

Die ALO-Library ist der Kern des gesamten Systems. Sie speichert und verwaltet die elektronischen Bücher. Die Library wird in Java implementiert und kann daher plattformübergreifend eingesetzt werden. Weiters gibt es für Java eine Vielzahl von freien Tools für die Verarbeitung von XML-Dateien und mittels *Servlets*⁴ eine Schnittstelle zu vielen bekannten Web-Servern.

4 Servlets sind Java Programme die vom Web-Server verwendet werden, um dynamische Inhalte zu erzeugen.

Die Bibliothek besitzt folgende Funktionen:

- Hinzufügen eines Buches in die Bibliothek
- Löschen eines Buches
- Ausgabe einer Liste der Bücher mit ihren Metadaten
- Ausgabe des Faksimile einer Buchseite
- Ausgabe des Texts einer Buchseite
- Ausgabe des gesamten Texts des Buches
- Ausgabe der Metadaten eines Buches

Der Buchtext wird in einer Datei im *TEI_xLite*-Format gespeichert. *TEI_xLite* ist ein XML-Format und hat daher den Vorteil, von Menschen und Computer lesbar sein. Das erhöht die „Haltbarkeit“ der Dateien und erleichtert die Publikation der Dateien im Internet. Beschreibende Metadaten wie Autor, Verlag, etc. können in die *TEI_xLite*-Datei gespeichert werden. Administrative und strukturelle Metadaten, die den Aufbau und die Formate des elektronischen Buches beschreiben, jedoch nicht. Diese Daten sind aber unbedingt notwendig um die „Haltbarkeit“ der Daten zu gewährleisten. Für diese Daten wird eine eigene Datei im Dateiformat *MOAII*, die für das Projekt „Making Of Amerika II“⁵ <<http://sunsite.berkeley.edu/moa2>> entwickelt wurde, verwendet. Das *MOAII*-Format ist wie *TEI_xLite* in XML definiert.

Die Faksimiles werden im GIF-Format abgespeichert. Das GIF-Format wurde gewählt, weil Dateien in diesem Format von allen Web-Browsern angezeigt werden können. Da die Faksimiles in Schwarz/Weiß vorliegen, kommt das Graphikformat JPEG nicht in Frage.

3.4 Management System

Das Management-System dient zur Erzeugung und Speicherung elektronischer Bücher. Es besteht aus einem Programm zur Erzeugung der *TEI_xLite*-Datei (*html2tei*), der *MOAII*-Datei (*makemoa*), zum Speichern des Buches in der *ALO*-Library (*addbook*) und zum Löschen eines Buches (*delbook*).

html2tei liest das von Finereader erzeugte HTML-File und wandelt sie in eine *TEI*-Datei um. Die beschreibenden Metadaten des Buches müssen von Hand in die *TEI*-Datei eingetragen werden.

makemoa erzeugt aus einer *TEI*-Datei und den Faksimile automatisch eine *MOA*-Datei für ein Buch.

5 Making of Amerika II beschäftigt sich mit der Erzeugung und dauerhaften Speicherung von digitalen Dokumenten

addbook speichert das aus TEI, MOA und Faksimile-Dateien bestehende elektronische Buch in der ALO-Library.

delbook löscht ein Buch aus der ALO-Library.

3.5 Benutzer-Interface

Das Benutzer-Interface der ALO-Library besteht aus statischen und dynamischen HTML-Seiten. *Abbildung 2* zeigt die vorhandenen HTML-Seiten und deren Verbindungen untereinander.

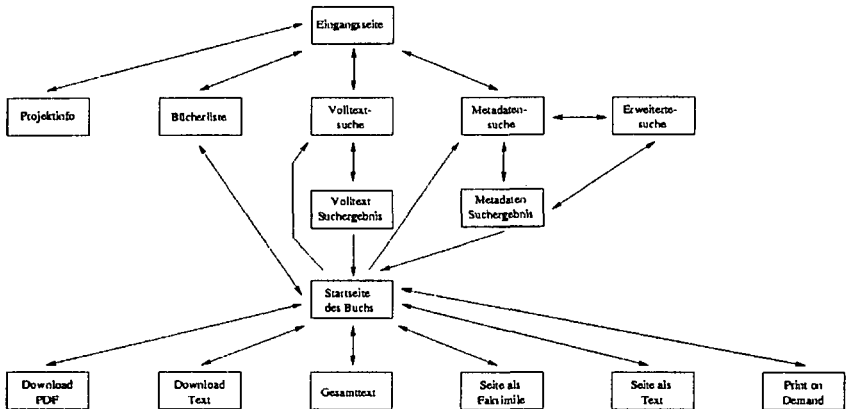


Abbildung 2: Verlinkung der ALO-Seiten

Seiten, die Buchinhalte darstellen, werden von Java-Servlets dynamisch erzeugt. Die Servlets stellen Anfragen an die ALO-Library und wandeln die XML-Antwort der Library mittels XSL-Stylesheets in HTML-Seiten um. Diese Vorgehensweise erlaubt eine schnelle und einfache Änderung der Darstellung durch das Ändern der Stylesheets. Dateien für den Download als PDF oder als Text werden auf die selbe Art erzeugt. In diesem Fall wird mit einem Stylesheet anstatt der HTML-Datei eine PDF, Text oder sonstige Datei erzeugt.

Die Metadatensuche ist ebenfalls ein Servlet, das in der Bücherliste der Library nach bestimmten Metadaten wie Autor, Verlag etc. sucht und danach eine Liste der gefundenen Bücher anzeigt. Für die Volltextsuche wird ein bereits vorhandenes System verwendet, das aus den TEI-Dateien einen Index erstellt und bei Anfragen in diesem Index sucht. Als Ergebnis einer Anfrage liefert das Programm eine Liste der Bücher, die die gesuchten Worte enthalten. Der Be-

nutzer kann daraus ein Buch auswählen und sich wiederum eine Liste der Seiten anzeigen lassen, in denen die Wörter enthalten sind.

Als BookOnDemand-Service wurde für das ALO-Projekt die Firma *Libri* <<http://www.libri.de>> gewählt. Einige ausgewählte Bücher der ALO-Bibliothek werden von Libri auf Knopfdruck gedruckt und geliefert. Die Bücher sind auch regulär über den Buchhandel bestellbar.

Literatur

- [Ebe00] Adolf Ebeling. Lesestunden Fünf OCR Klassiker im Vergleich. - In: c't magazin für computertechnik. Ausgabe 4/2000
- [MPW99] Alan Morrison, Michael Popham, and Karen Wikeander. Creating and Dokumenting Electronic Texts, A Guide to Good Practice. Oxford Books, 1999 Online <<http://ota.ahds.ac.uk/documents/creating/>>.

