Das eGovernment-Archiv der Virtuellen Fachbibliothek Ost- und Südostasien, CrossAsia

Anne Barckow, Matthias Gerhardt

Ostasienabteilung
Staatsbibliothek zu Berlin – Preußischer Kulturbesitz
Potsdamer Str. 33
10785 Berlin
x-asia@sbb.spk-berlin.de

Abstract: CrossAsia hat sich als erste Virtuelle Fachbibliothek das Archivieren amtlichen und halbamtlichen elektronischen Schrifttums zur Aufgabe gemacht. Ziel des Projektes ist die Erschließung, langfristige Sicherung und Bereitstellung der im Internet veröffentlichten Amtsdruckschriften aus Japan, Südkorea, der Volksrepublik China (inkl. Hongkong und Macao), Taiwan und Singapur und die Schaffung eines zentralen Zugangs zu den verteilten konventionellen und digitalen Materialien. Als Crawling-Software kommt Heritrix (Version 1.12.1) zum Einsatz, für die Indizierung wird NUTCHWAX in der Version v0.10.0 genutzt und für die Darstellung WAYBACK (Version v0.8.0).

1 CrossAsia, Virtuelle Fachbibliothek Ost- und Südostasien

CrossAsia[1], die Virtuelle Fachbibliothek Ost- und Südostasien, wird seit April 2005 unter Federführung der Staatsbibliothek zu Berlin in Zusammenarbeit mit acht nationalen und internationalen Partnern aufgebaut. Mit CrossAsia wird das im Rahmen des DFG-Förderprogramms „Überregionale Literaturversorgung“[2] an der Staatsbibliothek zu Berlin angesiedelte Sondersammelgebiet Ost- und Südostasien (SSG 6,25)[3] erweitert. Die Förderung der Virtuellen Fachbibliothek erfolgt ebenfalls durch die Deutsche Forschungsgemeinschaft (DFG). Seit dem Online-Gang von CrossAsia im März 2006 haben sich bereits mehr als 1250 Nutzer registriert. Das Portal bildet einen zentralen Einstieg auf der Suche nach ost- und südostasienrelevanten Ressourcen.

[1] http://crossasia.org

[2] http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/projektfoerderung/foerderziele/literaturerwerbung.html

[3] http://webis.sub.uni-hamburg.de/ssg/bib.1a/ssg.6_25

Kernmodule von CrossAsia sind die Metasuche über 16 Kataloge und Datenbanken, der Fachinformationsführer OGEA (Online Guide East Asia) als Nachweis- und Rechercheinstrument sämtlicher elektronischer Ressourcen der Ostasienabteilung der Staatsbibliothek zu Berlin und die Online-Contents-Datenbanken. Hier werden Aufsätze aus mehr als dreihundert westlichsprachigen Zeitschriften mit Bezug zu Ost- und Südostasien erfasst. Darüber hinaus bietet CrossAsia als einzige Virtuelle Fachbibliothek den überregionalen Zugriff auf mehr als vierzig lizenz- und registrierungspflichtige Datenbanken aus Ostasien. Neuestes Element von CrossAsia ist das eGovernment-Archiv. Webseiten asiatischer Regierungsinstitutionen inklusive der darauf hinterlegten Volltexte werden mit dem Einverständnis der Rechteinhaber in regelmäßigen Intervallen gespeichert und über die Seiten von CrossAsia zur Verfügung gestellt.

2 eGovernment-Archiv CrossAsia

Ziel des Aufbaus eines Archivservers für die Virtuelle Fachbibliothek ist, unter Einsatz Heritrix (Version 1.12.1) als Crawling-Software, NUTCHWAX in der Version v0.10.0 für die Indizierung und WAYBACK (Version v0.8.0) für die Darstellung ein vielschichtiges System für die Indizierung und Suche zu entwickeln, das den Erfordernissen, Online-Ressourcen lokal vorzuhalten, entspricht. Eine Integration in größere Nutch-Instanzen ist problemlos möglich.

Gleichzeitig bedeutet der Aufbau eines eGoverment-Archivs eine Erweiterung des an der Staatsbibliothek zu Berlin angesiedelten Sammelschwerpunktes Parlamentsschriften. Die Verantwortung für den regionalen Ausschnitt Ost- und Südostasien liegt bei der Ostasienabteilung der Staatsbibliothek zu Berlin. Publikationen amtlicher Stellen sind eine für das gesamte Fächerspektrum der Asienwissenschaften unersetzliche Quelle. Der Umfang ist beachtlich. Amtliche Literatur ist oft nicht über den Buchhandel erhältlich, und auch die Nachweissituation von Print-Ausgaben genügte trotz Pflichtexemplargesetz in den asiatischen Ländern häufig nicht den Erfordernissen der Wissenschaft.

Vor allem in Japan gibt es seit Beginn des Jahrtausends Initiativen, die Ausgangslage zu verbessern. Dazu zählt das „e-Japan戦略“-Programm der Regierung und seine Erweiterungen[4] sowie WARP (Web ARchiving Project)[5], realisiert von der National Diet Library. In anderen Regionen lassen sich ähnliche Programme nicht bzw. nicht im selben Ausmaß und in absehbarer Zeit erwarten, und auch in Japan werden Dokumente z.B. von Regierungsinstitutionen angegliederten Einrichtungen noch nicht in der aus Sicht der Ostasienwissenschaften erforderlichen Tiefe erschlossen.

Aufbauend auf die lokalen Projekte setzt CrossAsia hier mit dem Aufbau des Digitalen Archivs an. Ziel des Projektes ist die Erschließung und Sicherung der im Internet

[4] Links zu den Volltexten des Programms unter http://www.kantei.go.jp/jp/singi/it2/ (Japanisch) oder http://www.kantei.go.jp/foreign/policy/it/index_e.html (Englisch)
[5] http://warp.ndl.go.jp/

verfügbaren ostasiatischen Amtsdruckschriften und die Schaffung eines zentralen Zugangs zu den verteilten konventionellen und digitalen Materialien.

Im Rahmen des Projekts wurden alle Tools konfiguriert und getestet. Die für die Archivierung vorgesehenen Ressourcen sind selektiert und mit Metadaten erschlossen. Derzeit wird die Genehmigung der Rechteinhaber zur Archivierung und Bereitstellung auch älterer Versionen der Dokumente eingeholt. Bis zur Informatik 2008 wird das eGovernment-Archiv[6] in die Virtuelle Fachbibliothek Ost- und Südostasien eingebunden.

2.1 Inhalt

Das eGovernment-Archiv von CrossAsia bezieht sich regional auf die VR China (inkl. Hong Kong und Macao), Taiwan, Singapur, Japan und Korea. Akzente sollen in Bereichen gesetzt werden, wo ein dauerhafter Zugriff auf wissenschaftlich relevante Ressourcen nicht gewährleistet scheint oder die Entwicklungen und ihre Rezeption im WWW besonders rasant und unübersichtlich sind.

Im eGovernment-Archiv von CrossAsia werden die Webseiten amtlicher und halbamtlicher Stellen aus der Region und die darauf hinterlegten Veröffentlichungen gesammelt. Gegenstand des Archivs sind Statistiken, Weißbücher, Kommissionsberichte, Auftragsstudien, Strategiepapiere, Programme etc. Dabei ist Ziel die möglichst weitgehende Erfassung von Amtsschriften im Sinne der von der Official Publications Section der IFLA im August 1983 veröffentlichten Definition von Amtlichen Veröffentlichungen für den Internationalen Gebrauch.

2.2 Technische Umsetzung

Seit 2003 beschäftigt sich das International Internet Preservation Consortium[7] mit der Entwicklung von open source Programmen für Download, Archivierung und Zugriff von Webseiten. Diese sind seit 2005 für die allgemeine Nutzung zugänglich. Da diese Werkzeuge international von Staatsbibliotheken zur Archivierung des nationalen Webs verwendet werden, passt sich CrossAsia diesen Standards an. Als Crawling-Software kommt Heritrix (Version 1.12.1) zum Einsatz, NUTCHWAX in der Version v0.10.0 für die Indizierung und WAYBACK (Version v0.8.0) für die Darstellung. Damit werden folgende Zielsetzungen erfüllt: Erreichung internationaler Standards, Kompatibilität und Vernetzbarkeit mit anderen Projekten, leichte Archivierung, da Heritrix die Formatvielfalt sofort auf ein einheitliches Format reduziert, und Volltextsuchbarkeit.

[6] http://crossasia.org/de/archive/
[7] http://netpreserve.org/about/index.php

Alle Tools wurden für die lokalen Bedürfnisse konfiguriert und aufeinander abgestimmt. Von besonderer Bedeutung ist dabei die Berücksichtigung der verschiedenen Formate und besonders Kodierungen. Das eGovernment-Archiv von CrossAsia selbst und die Metadaten sind UTF-8 kodiert. Die zu archivierenden Ressourcen allerdings weisen sehr heterogene Kodierungen auf: neben UTF-8 finden sich Dokumente in GB 2312 (Volksrepublik China), BIG-5 (Taiwan), Shift-JIS, EUC-JP (Japan) oder EUC-KR (Korea) etc.[8] [Lu00] Die Suchmaschine des Archivs muss unterschiedlich kodierte Dokumente (unterschiedliche Zeichensätze) und darüber hinaus mehrere Varianten eines Zeichens wie Lang- und Kurzzeichen mit unterschiedlichen (Unicode-)Werten simultan recherchierbar anbieten. Deswegen ist die Eingabe in die Suchmaske Unicode / UTF-8 kodiert; ein Filter hinter der Eingabemaske wird die Umkodierungsmöglichkeiten auf Grundlage sehr komplexer Tabellen realisieren.

Eingebunden ist das Archiv in den Online Guide East Asia (OGEA), das Erfassungs- und Nachweisinstrument für alle elektronischen Ressourcen der Ostasienabteilung. OGEA wurde im Rahmen von CrossAsia eigenentwickelt. Das Frontend basiert auf der Programmiersprache PHP und AJAX-Technologien. Das Backend basiert auf dem relationalen Datenbanksystem MySQL. Alle Daten werden in UTF-8 Format in der Datenbank vorgehalten- Damit wird der Mehrschriftlichkeit der zu erfassenden Daten Rechnung getragen.

Ein integrierter Workflow zur Archivierung in der Erfassung ermöglicht es dem Erfasser der Ressource schnell und unkompliziert den Datensatz für das Archiv vorzubereiten. Dabei muss das Archivierungsintervall gesetzt sein. Über automatisierte Prozesse werden die Daten aus der MySQL-Datenbank gefiltert und in die vorbereiteten Profile in Heritrix eingebunden. Heritrix selbst wird über „Cron-Jobs“ gestartet, welche die verschiedenen Intervalle ansprechen. Nach erfolgreichem Crawl der Internetpräsenzen wird mit Hilfe von NutchWax, welches nur in Zusammenarbeit mit Hadoop – einem kleinen Teil der Lucene-Such-Engine – die als .arc archivierten Daten aus Heritrix indexiert. Um die Anzeige der archivierten Daten kümmert sich die Wayback-Machine.

Die drei Applikationen sind reine JAVA-Anwendungen und benötigen den Tomcat-Apache.

Die folgende Abbildung verdeutlicht noch einmal den Prozess der Archivierung und Darstellung von elektronischen Ressourcen in der virtuellen Fachbibliothek Ost- und Südostasien – CrossAsia.

[8] Zu den gebräuchlichen ostasiatischen Kodierungen weltweit s. Lunde, Ken: CJKV information processing : [Chinese, Japanese, Korean & Vietnamese computing]. Beijing [u.a.], O'Reilly: 1999.

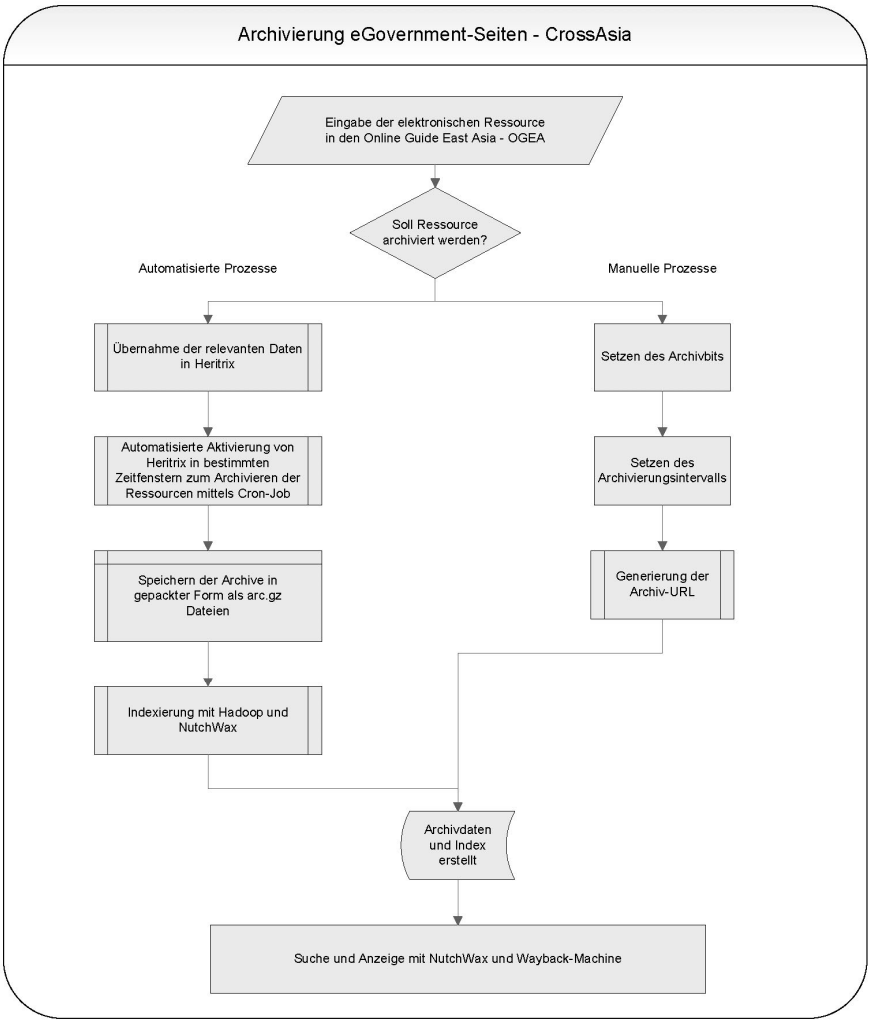

Abbildung 1: Archivierung eGovernmentseiten CrossAsia

Die Metadaten zu den einzelnen Ressourcen sind über CrossAsias Online Guide East Asia[9] recherchierbar. Die Volltexte der archivierten Versionen lassen sich über das eGovernment-Archiv in einfacher oder erweiterter Suche sowie über eine URL- und Department-List recherchieren.

[9] http://crossasia.org/de/ogea/

2.3 Workflow

Zunächst ist eine Übersicht der zahlreichen Regierungswebseiten Japans, der Volksrepublik China, Hong Kongs, Macaos, Singapurs und Taiwans sowie eine Downloadstrategie angefertigt worden. Parallel wurden in OGEA, als Erfassungsintrument für die elektronischen Ressourcen der Ostasienabteilung Archivierungsroutinen implementiert.

In OGEA, dessen Metadatenschema sich nach den Vorgaben von Dublin Core und dem vascoda Application Profile richtet, werden die Metadaten der für das Archiv ausgewählten Quellen arbeitsteilig mit dem Seminar für Japanologie der Universität Tübingen, dem Seminar für Sinologie und Koreanistik der Universität Tübingen und dem Institut für Sinologie der Universität Heidelberg erfasst und formal und inhaltlich umfassend beschrieben. Mit der Zuordnung eines Metadatensatzes zum eGovernment-Archiv wird über ein in die Datenbank implementiertes Kontaktformular beim Rechteinhaber die Genehmigung zum Speichern und Bereitstellen der Ressourcen eingeholt. Dieser Vorgang wird nach 30 Tagen wiederholt, falls keine Reaktion erfolgt ist. Die Korrespondenz wird dokumentiert und archiviert.

Für zu archivierende Quellen wird eine Archiv-URL generiert und die Archivierung gestartet. Ein Abzug der Ressource wird regelmäßig im definierten Intervall gemacht.

3 Ausblick

Bis zum Sommer 2008 wird das eGovernment-Archiv über die Seiten von CrossAsia öffentlich zugänglich gemacht. Mittels Open-URL wird das Archiv in die Metasuche von CrossAsia eingebunden werden. Auf technischer Ebene geht die Erweiterung darüber hinaus dahin, Metadaten und Volltexte zu den Dokumenten gemeinsam durchsuchbar zu machen. Die Suchfunktionalitäten sollen ausgeweitet werden.

Inhaltlich wird der nächste Schritt dahin gehen, die Archivierungstätigkeiten von Institutionen auf staatlicher Ebene auf weitere Verwaltungsebenen (Präfekturen, Gemeinden etc.) auszudehnen Grundlage für eine Entscheidung für zu archivierende Inhalte ist das Erwerbungsprofil der Ostasienabteilung der Staatsbibliothek zu Berlin. Ein Schwerpunkt wird damit auf die Geistes- und Sozialwissenschaften gelegt werden.

Das eGovernment-Archiv wird Bestandteil eines vernetzten, fachübergreifenden Repositoriums an der Staatsbibliothek zu Berlin werden. Somit ist eine gemeinsame Recherche auch mit anderen elektronischen Volltexten wie OpenAccess-Zeitschriften und weiteren auf den Servern der Staatsbibliothek zu Berlin hinterlegten Dokumenten möglich.

Literaturverzeichnis

[Lu00] Lunde, Ken: CJKV information processing : [Chinese, Japanese, Korean & Vietnamese computing]. O'Reilly, Beijing [u.a.] 1999.