

Archivierung von landeskundlichen Netzpublikationen

Ein Projekt der Rheinischen Landesbibliothek und des Hochschulbibliothekszentrum Köln

1. Idee

Der im Mai 2003 begangene zehnte Geburtstag des Internets verdeutlicht, wie grundlegend sich seitdem nicht nur die Bereitstellung von Information verändert und erleichtert hat. Die Zugriffsmöglichkeiten zu jeder Zeit und von fast allen Orten der Welt haben eine neue Form von Öffentlichkeit entstehen lassen. Im Gefolge unterliegt auch das Publikationsverfahren vielfältigen Umbrüchen, die den neuen Möglichkeiten Rechnung tragen. Wenngleich weiterhin ein Großteil der Veröffentlichungen gedruckt erscheint, sind zum Teil bereits heute manche Informationen nur in elektronischer Form greifbar, weil für weitgehend aktuelle Inhalte oder Darstellungen mit begrenztem Interessentenkreis das elektronische Publizieren zu einer kostengünstigen und unkomplizierten Alternative geworden ist. Gerade Websites bieten Chancen der Selbstdarstellung, die bis vor kurzem undenkbar waren.

Wollen sich Bibliotheken nicht diesem expandierenden Informationssegment verschließen und damit langfristig Informationskompetenz einbüßen, müssen sie diese Herausforderung annehmen.

Vorreiter auf diesem Gebiet waren nicht zufällig die Universitätsbibliotheken, da an den Universitäten und anderen Hochschulen schon bald Hochschulschriften elektronisch publiziert wurden, deren Bereitstellung und dauerhafte Archivierung die einzelnen Institute und Fachbereiche teilweise überforderte (vgl. das Projekt DissOnline¹). Als Die Deutsche Bibliothek (DDb) im Frühjahr 2002 bekanntgab, sich auf die Sammlung von E-Publikationen des Verlagsbuchhandels² konzentrieren zu wollen, drohten große Teile der elektronischen Veröffentlichungen aus dem Blick bibliothekarischer Sammeltätigkeit zu geraten. Zwar besteht in dem „public non-profit“ Internet Archive³ (San Francisco, Ca.) die Möglichkeit, mittels einer wayback machine sich die Schnappschüsse von Homepages chronologisch geordnet anzeigen zu lassen. Jedoch ist weder das Einsammeln einer Website gewährleistet noch ein nachvollziehbarer Sammelrhythmus zu erkennen. Außer der Suche nach einer URL bietet sich dem Nutzer keine weitere sachliche oder formale Erschließung.⁴ Hier begann man in der Rheinischen Landesbibliothek Koblenz mit Überlegungen, elektronische Publikationen außerhalb des Buchhandels zu sammeln, sofern sie von Anbietern aus dem Pflichtexemplarbereich stammen oder von landeskundlichem Interesse sind. Somit weitete sie den ohnehin bestehenden Sammelauftrag der Landesbibliothek auf Netzpublikationen aus.

Das Pflichtexemplarrecht sowohl des Bundes als auch der Länder sieht jedoch zur Zeit noch keine Regelung für die Abgabe und Archivierung elektronischer Dokumente vor. Deshalb streben Die Deutsche Bibliothek (für den Bund) gemeinsam mit den Regionalbibliotheken sowie den zuständigen Ministerien (für die Länder) eine gesetzliche Neuregelung an. Bis diese Novellierung in Kraft treten wird, schränkt die geltende Rechtslage des Pflichtexemplar-

¹ <http://www.dissonline.de/>

² Ute Schwens: Die Deutsche Bibliothek – gesetzlicher Auftrag und elektronische Publikationen. In: ZfBB 49 (2002) 1, S. 13-17, 15.

³ <http://www.archive.org/>

⁴ Vgl. etwa: Werner Pluta: Digitales Alexandria. Archive sichern Internet-Schätze. In: c't (2003), H.11, S.172-175, 174.

und Urheberrechts das Einsammeln, Archivieren und Bereitstellen von elektronischen Dokumenten ein.

So scheiterte der ursprüngliche Plan, mittels eines „klug programmierten Harvesters“ (harvester = Erntemaschine, Sammler) automatisch alle betreffenden Websites einzusammeln und zu archivieren, nicht nur an technischen Problemen⁵, sondern auch an der gegenwärtigen Urheberrechtssituation. Das Spiegeln einer Website auf einem anderen, öffentlich zugänglichen Server erfüllt nach der Rechtslage den Tatbestand der Verbreitung und bedarf daher einer Genehmigung durch den Rechteinhaber.

Ähnliche Probleme gibt es wohl auch in anderen Ländern. So sind zum Beispiel die eingesammelten Websites des Nordic Web Archive⁶, einer Initiative der Nationalbibliotheken Dänemarks, Finnlands, Schwedens, Norwegens und Islands, über das Internet nicht einsehbar. Statt dessen besann man sich in Koblenz der Kernqualitäten einer Landesbibliothek: Informations- und Bewertungskompetenz in regionalen Zusammenhängen. Inspiriert von dem PANDORA-Projekt⁷ der National Library of Australia rückte das Ziel einer inhaltlich bewerteten Auswahl von E-Publikationen und Websites in den Blick.

2. Vorgehensweise und Tests

Im Spätsommer 2002 ergab sich bei einem Kundengespräch zwischen dem Hochschulbibliothekszentrum Köln (HBZ) und der Rheinischen Landesbibliothek die Möglichkeit, ein gemeinsames Projekt zur Sammlung, Bereitstellung und Dauerarchivierung von Websites⁸ und elektronischen Pflichtexemplaren aufzubauen.

Das HBZ sammelte zu diesem Zeitpunkt erste Erfahrungen als Dokumentenserver-Host für einige nordrhein-westfälische Fachhochschulen.

Nach eingehender Prüfung fiel die Entscheidung zugunsten des an der UB Stuttgart entwickelten Dokumentenverwaltungssystem OPUS (Online Publikationsverbund der Region Stuttgart). Die ausschlaggebenden Pluspunkte waren:

- modularer Aufbau
- gute Performance
- individuelle Anpassungen möglich
- in Deutschland an der UB Stuttgart entwickelt
- Weiterentwicklung ist gewährleistet
- vielfältig eingesetzt in deutschen Bibliotheken
- nicht kommerzielle Software
- verfügt über OAI-Schnittstelle

Da sich der technische Sammelvorgang bei verschiedenen Dokumentarten unterschiedlich gestaltet, wurden zwei Server aufgebaut:

ein Server, der WWW-Seiten einsammelt, und ein Dokumentenserver, der mit elektronischem Material bestückt wird, das formal weitgehend den konventionellen Publikationsformen entspricht, die auch gedruckt gesammelt worden wären, insbesondere Zeitschriften, Hochschul- und Amtsdruckschriften sowie seit neuestem Firmen- und Vereinsschrifttum.

⁵ Projekte wie das Nordic Web Archive der skandinavischen Nationalbibliotheken nährten solche Vorstellungen vgl. Nikola Korb, Berthold Weiß: The Nordic Web Archive. In: Dialog mit Bibliotheken 14 (2002) 1, S.30-32.

⁶ Svein Arne Brygfjeld: Access to Web archives: The Nordic Web Archive Access Project approach. In: ZfBB 49 (2002), S.227-231.

⁷ <http://pandora.nla.gov.au/index.html>

⁸ Der Begriff „Website“ wird hier verwandt wie er in: *Der Brockhaus Computer und Informationstechnologie.* – Mannheim: Brockhaus, 2003 (S. 813 und 968) definiert wird.

2.1. RLB-Webarchiv

Von der Idee des vollautomatischen Einsammelns von Websites wurde - wie oben bereits erwähnt - schnell Abstand genommen. Um einen Harvester so zu programmieren, dass er relevante Dokumente vollautomatisch einsammelt, bedarf es einiger Forschungsarbeit. Die in Skandinavien angewendete Methode, die Länderkennung der URL als Selektionszeichen zu nutzen, konnte bei einer Sammlung, die auf ein deutsches Bundesland beschränkt bleibt, nicht funktionieren. Darüber hinaus existieren landeskundlich relevante Sites, deren URL beispielsweise auf .org oder .com enden, von ausländischen Angeboten ganz zu schweigen.

Deshalb entschieden sich die Projektpartner, erste Erfahrungen mit dem Einsammeln einzelner Websites zu machen und diese sachlich geordnet anzubieten, um einen Mehrwert gegenüber den oben erwähnten Wayback-Maschinen zu erzielen.

In ersten Tests wurden die Websites halbautomatisch mit Hilfe der Programms w3mir⁹ eingesammelt und diese Daten gezippt zur Verfügung gestellt.

w3mir erwies sich schnell als unpraktisch, da diese Software viel Arbeitsspeicher benötigte, zu langsam arbeitete und das Sammelergebnis unbefriedigend ausfiel. Das Zippen der eingesammelten Dateien war zudem nutzungsunfreundlich.

Deshalb fiel die Wahl auf den Open Source Offline Browser httrack¹⁰, der auch im Pandora-Projekt der National Library of Australia eingesetzt wird.

Sammelvorgang und Bearbeitung

Nach Eingabe der URL in ein Formular wird per Knopfdruck die gewünschte Website eingesammelt. Es besteht die Möglichkeit, die Sammelfunktion auf bestimmte Verzeichnisse, Dateiarten oder auf bestimmte Hierarchiestufen zu beschränken. Bei passwortgeschützten Seiten kann dem Gatherer (to gather = ernten, pflücken, raffen) die Login-Kennung und das Passwort mitgegeben werden.

⁹ <http://langfeldt.net/w3mir/>

¹⁰ <http://www.httrack.com>

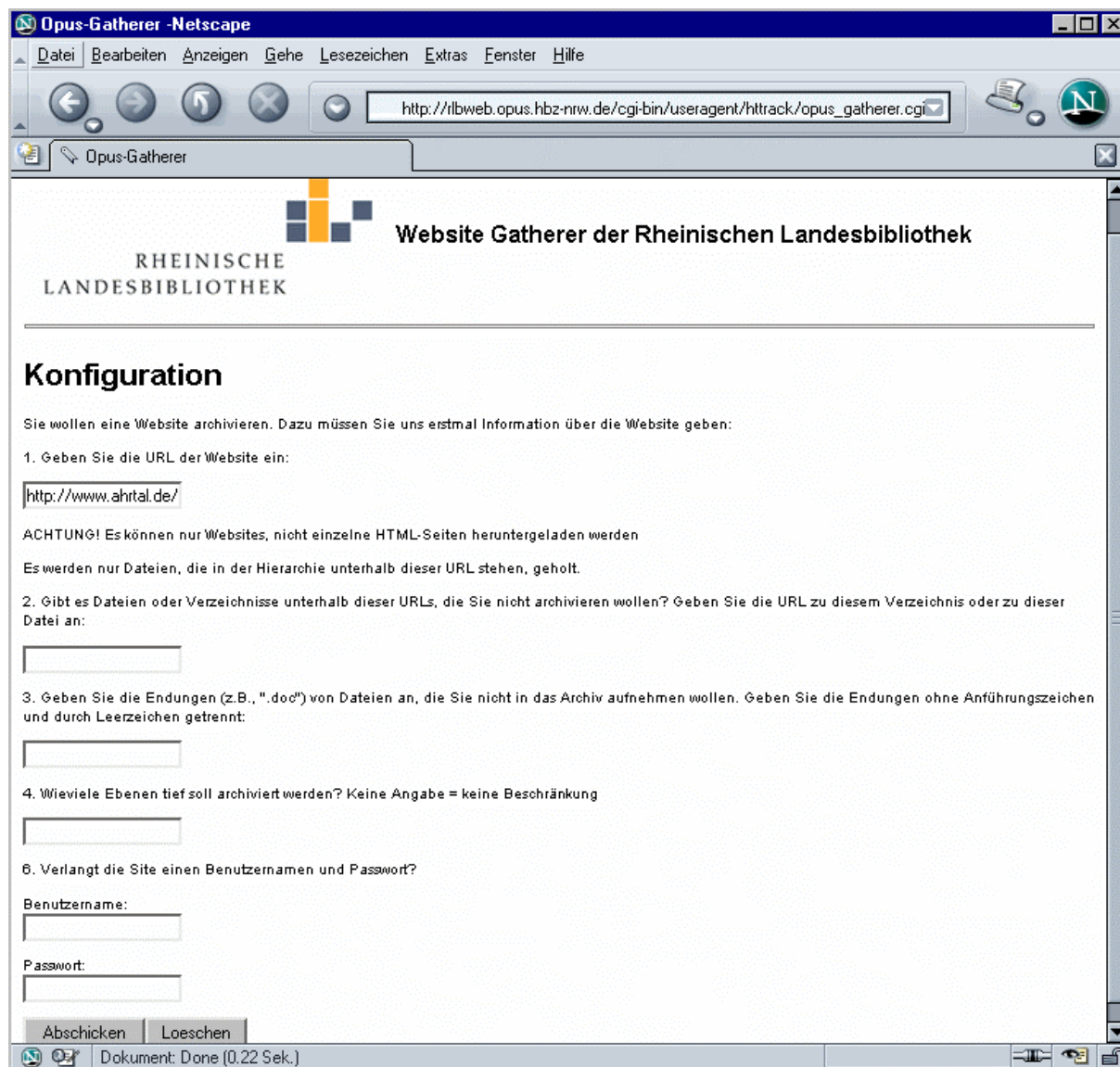


Abb. 1: Eingabeformular des Gatherers, in dem die Optionen für das Einsammeln der Webseiten festgelegt werden

Die zu dieser Website gehörenden Dateien werden im Browser in Listenform angezeigt. Am Ende der Liste hat man die Möglichkeit, per Mausklick wieder zurück zum Gatherer zu wechseln, um die Konfiguration für den Sammelvorgang zu ändern und diesen neu zu starten, wenn das Ergebnis nicht zufriedenstellend war, oder die eingesammelten Daten zu übernehmen.

Danach werden per Mausklick die dem Dokument mitgegebenen Metadaten aus dem HTML-Title-Tag sowie aus eventuell vorhandenen Meta-Tags¹¹ per Skript herausgezogen und in die Metadatenfelder der OPUS-Datenbank übertragen (z.B. Title-Tag in Kategorie „Titel“ oder die Daten aus dem Meta-Tag „keywords“ in das Feld „freie Schlagwörter“).

Im sogenannten Administrationsmodul von OPUS ist eine Korrektur und Ergänzung der eingesammelten Metadaten möglich. Abschließend werden sie von einem provisorischen in den endgültigen Status gehoben und indexiert, damit sie recherchierbar sind. Diese Konstruktion ermöglicht zum Beispiel das Melden einer Website durch den Inhaber oder Ersteller der Site; die endgültige Aufnahme bleibt aber der Bibliothek vorbehalten. Ein Löschen der Daten sowie der dazugehörigen Dokumente ist im Administrationsmodul ebenso möglich.

¹¹ Ein Meta-Tag ist ein HTML-Kommando im Kopf einer HTML-Datei, das mit dem Tag (dt. Marke) <META> eingeleitet wird. Darin können Informationen über das Dokument wie Inhaltsangabe, Suchbegriffe, etc. enthalten sein, die u.a. von Suchmaschinen ausgewertet werden können.



Ändern eines Eintrags in OPUS

subject_type2 ist geo

Zusätzliches Autorenfeld anfordern

Originaltitel:

Englischer Titel:

Deutscher Titel:

1. Verfasser:

Urheber:

Sonst. Beteiligter:

Sonst. bet. Inst.:

Schlagwörter SMD: [SMD](#)

SW unkontr. deutsch:

SW unkontr. englisch:

Abstract Originalsprache:

Das Ahrtal im Internet. Internetmagazin rund um das Ahrtal. Inhalt: Freizeitangebote, Kunst & Kultur, Weinkeller & Weinlokale, Skurriles & Witziges, Hotels, Veranstaltungskalender, Ferienwohnungen, Shareware, Diashow usw. usw.. Hier finden Sie einfach alles über das Ahrtal.

Abstract in einer weiteren Sprache:

Erstellungsjahr:

Dokumentart:

DNB-Sachgruppe:

Sprache:

Quelle:

Identifizier:

email:

Gültig bis: (Format 31.01.2000, 0 für unbeschränkt)

RPB - Systematik:

Geografische Systematik:

Volltext: [26-index.html](#)
Publikationsdatum: 18.02.2003
RLB-Webarchiv-Ident-Nr.: 26

Abb. 3: Metadaten-Formular im Administrationsmodul

Um den Arbeitsaufwand zu verringern, wurde die Nachbearbeitung der Metadaten, bis entsprechende Standards vereinbart sind, darauf beschränkt, den Hauptsachtitel nach Vorlageform aufzunehmen und die herausgebende Körperschaft nach GKD anzusetzen. So erschien bei den automatisch gewonnenen Metadaten beispielsweise anstelle des Namens der Gemeinde, für welche die Homepage erstellt wurde, im Titelfeld lediglich ein unverständliches „@edomain“. Häufig wurde in das Verfasserfeld der Webdesigner automatisch übernommen, der nachträglich vom Bearbeiter in das Feld „Sonstiger Beteiligter“ transferiert wurde. Ob und in welchem Maße eine nachträgliche Bearbeitung der Daten zu leisten ist, wird voraussichtlich erst in Absprache mit DDB und den Verbänden zu regeln sein. Die geplante Volltextsuche erlaubt ganz andere Suchmöglichkeiten, die wahrscheinlich traditionelle Formen der bibliothekarischen Formal- und Sacherschließung nur noch bedingt erfordern.

Zusätzlich besteht optional die Gelegenheit, das normierte Schlagwortvokabular der Schlagwortnormdatei oder freie, das heißt unkontrollierte Schlagwörter zu vergeben. Obligatorisch ist hingegen die Erschließung des Dokuments durch zwei unterschiedliche Systematiken.

Auf Empfehlung der DINI-Arbeitsgruppe Elektronisches Publizieren soll eine inhaltliche Beschreibung erfolgen, die sich an den Sachgruppen der Deutschen Nationalbibliographie orientiert.¹² Ein Umstieg auf die Dewey Decimal Classification steht unmittelbar bevor. Die Ausrichtung des Sammelauftrags auf regionale und landeskundliche Dokumente erforderte die Erweiterung des OPUS-Schemas. Da seit 1991 alle Rheinland-Pfalz betreffenden Publikationen in der Rheinland-Pfälzischen Bibliographie (RPB) verzeichnet werden, lag es nahe, die zur Erschließung der konventionellen Medien verwendete Systematik auch für die Netzpublikationen zu verwenden. Momentan ist als drop-down-Menü eine Version der Systematik integriert, die lediglich die Hauptgruppen abbildet. Eine Erweiterung auf alle Sachstellen ist jedoch bereits in Arbeit.

Im Hinblick auf die mögliche Ausweitung des Modells auf sämtliche Pflichtexemplarbibliotheken wurde eine ausbaufähige geographische Systematik in das Metadatenschema aufgenommen, die eine Suche nach regionalen und geographischen Aspekten ermöglicht.

Abschließend werden die URN (Uniform Resource Number) berechnet und per Mausclick die Metadaten indexiert.

2.2. RLB-Dokumentenserver

Im Gegensatz zum Webserver werden auf dem Dokumentenserver Netzpublikationen konventioneller Publikationsart gesammelt.

Für das Erfassen der Dokumente gibt es ein eigenes Eingabeformat, welches je nach Dokumentart (Aufsatz, Dissertation, Report, ...) variiert und auf dem Dublin Core Metadatenformat beruht.

Im Gegensatz zu den Websites werden diese Dokumente wie die vergleichbaren gedruckten Publikationen formal und sachlich erschlossen. Ansonsten gleicht die Bearbeitung der Metadaten in etwa der beschriebenen Vorgehensweise bei Websites: auch hier wird die URN berechnet und die Datei bzw. die Dateien auf den Server geladen. Ebenso gibt es für die

¹² Inhaltliche Gestaltung der OAI-Schnittstelle. Eine Empfehlung für Daten-Provider an deutschen Universitäten, S.2., http://www.dini.de/documents/dini_oai_empfehlungen_07-2002.pdf

endgültige Übernahme in die Datenbank ein Administrationsmodul mit der Möglichkeit der Nachbearbeitung.

2.3. Suche in Web- und Dokumentenserver

Geboten werden zur Zeit drei unterschiedlich geartete Sucheinstiege.

Zum einen gibt es die konventionelle, bibliothekarische Form der Suche in den Metadaten mit den Abfragemöglichkeiten in verschiedenen Kategorien – vom Titelstichwort bis zu den unterschiedlichen Klassifikationen.

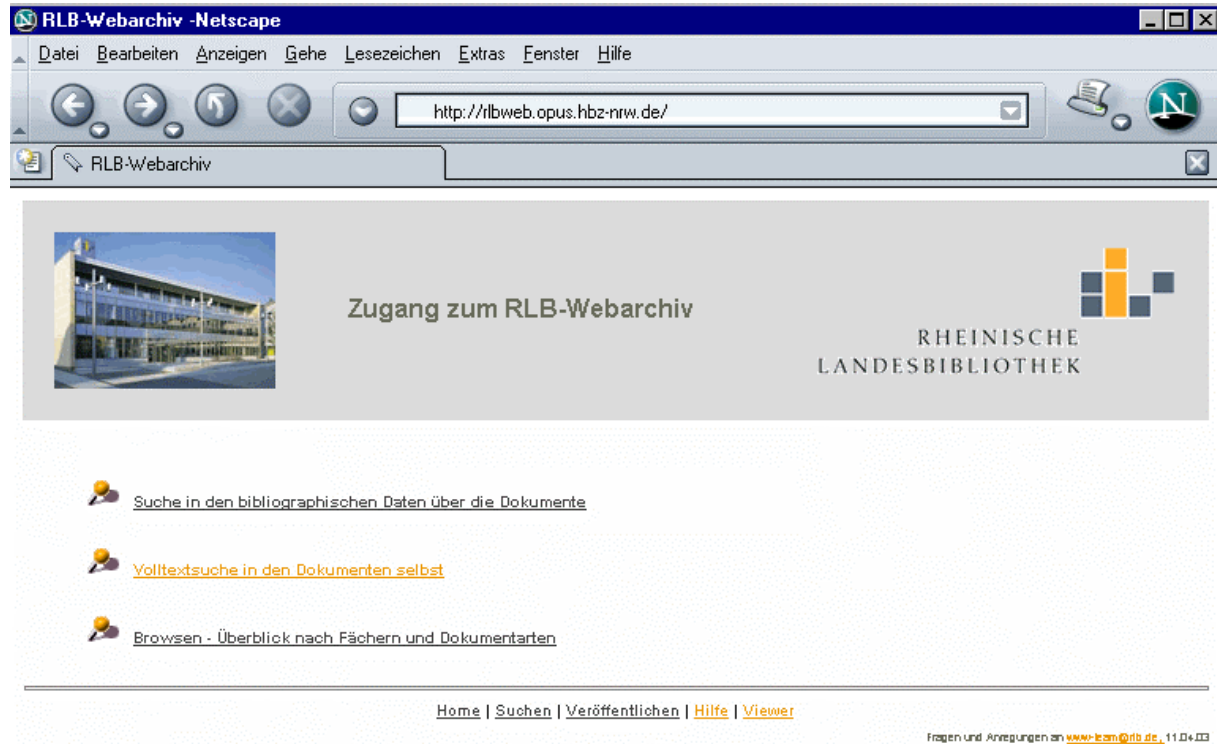


Abb. 4: Startseite des Webarchivs

Zum anderen ist eine Volltextsuche geplant, die es erlaubt, im kompletten Datenbestand der gesammelten Dokumente zu recherchieren.

Um einen sachlichen Zugriff auf die E-Publikationen zu bieten, ist eine Browsing-Funktion installiert, die wahlweise eine Suche in den Sachgruppen der Deutschen Nationalbibliographie (DNB), der Rheinland-Pfälzischen Bibliographie oder nach den geographischen Räumen ermöglicht.

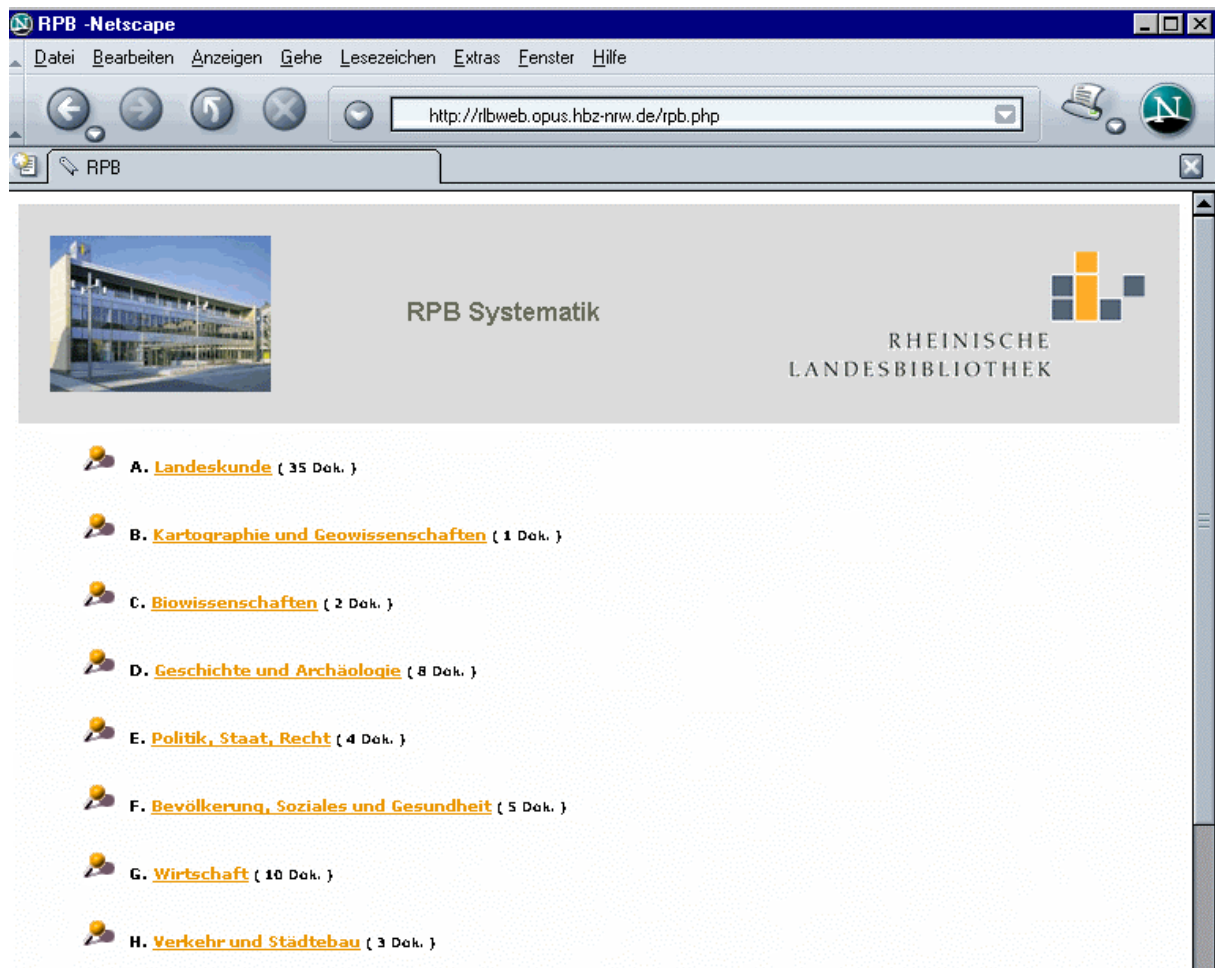


Abb. 5: Browsen in der Systematik der Rheinland-Pfälzischen Bibliographie (RPB)

Um den Nutzern größeren Komfort zu bieten, soll eine gemeinsame Suche über Web- und Dokumentenserver kurzfristig realisiert werden.

3. Projektergebnisse

Nach einer gut halbjährigen Testphase können die folgenden Erkenntnisse und Ergebnisse festgehalten werden:

- Die Aufteilung in zwei Server für Websites bzw. für („klassische“) elektronische Dokumente ist wegen der unterschiedlichen Handhabung beim Einsammeln notwendig. Entgegen ersten Vorstellungen wird es jedoch eine gemeinsame Suche sowie ein gemeinsames Browsen durch die Systematik der RPB und die Systematik der Deutschen Nationalbibliographie für beide Dokumentarten geben.
- Die URN, in Deutschland bislang nur für Dissertationen vergeben, wird erstmals auch für andere Dokumentarten genutzt. In enger Zusammenarbeit mit der URN-Stelle der Deutschen Bibliothek werden für die gesammelten Dokumente URNs vergeben und die dazugehörigen Daten an die DDB gemeldet.
- Es soll nicht verschwiegen werden, dass das Einsammeln von dynamisch erzeugten Websites und solchen Sites, die mit Hilfe eines Content Management Systems (CMS) erstellt werden, nicht unproblematisch ist. Der Gatherer ist teilweise nicht in der Lage,

einzelne Dateien der Website oder die Site als Ganzes einzusammeln. Hier besteht sicherlich noch Forschungsbedarf.

- OPUS entstammt als Dokumentenverwaltungssystem dem Hochschulbereich und ist auf dort auftretende Dokumentarten vorrangig monographischen Charakters zugeschnitten. Für den Regelbetrieb und insbesondere im Hinblick auf den geplanten Nachweis der Metadaten in den Verbundkatalogen oder der ZDB ist eine Erweiterung des Schemas für periodische Publikationen und damit verbundene hierarchische Strukturen erforderlich.
- Da die Sammeltätigkeit bislang ohne gesetzliche Grundlage erfolgt, wird bei jedem Rechteinhaber die Genehmigung eingeholt, die betreffenden Dateien auf den Servern zu spiegeln. Mittels einer kurzen Formmail wird der Adressat mit dem Anliegen bekanntgemacht. In einer im Anhang mitgesandten Erklärung gibt der Rechteinhaber der RLB das Einverständnis, seine Website weltweit in Datennetzen zur Verfügung zu stellen. Die bisher Angeschriebenen haben in den allermeisten Fällen positiv reagiert und überraschend schnell ihre Zustimmung gegeben.
- Ungeklärt ist bis heute die Frage, wie die eingesammelten Dokumente archiviert werden sollen. Die weiterhin rasante Entwicklung im IT-Bereich lässt voraussehen, dass Websites aus dem Jahr 2003 in einigen Jahren nicht mehr mit den dann gängigen Tools genutzt werden können. Hier müssen die nationalen und internationalen Entwicklungen verfolgt werden, da dieses Thema nur in diesem Kontext angegangen werden kann¹³. Die noch offene Frage der Dauerarchivierung und zukünftigen Nutzung sollte aber kein Vorwand sein, auf eine zumindest selektive Webarchivierung zu verzichten.
- Das australische PANDORA-Projekt beschränkt die Sammlung allein auf qualitativ hochwertige Seiten. Um zu dokumentieren, wie Rheinland-Pfalz, seine Institutionen und Menschen sich im Internet darstellen, greift diese Selektion unter allein qualitativen Aspekten indes zu kurz. Fast jede Gemeinde, jede Firma, aber auch jeder Verein präsentiert sich heute im Internet: Als Pendant zu den gedruckten Fest- und Jubiläumsschriften ist eine Selbstdarstellung im Web zumindest in exemplarischer Auswahl zu archivieren. Dokumente dieser Art könnten hohen Quellenwert für das frühe 21. Jahrhundert bekommen. Ob jedoch exakt gefasste Sammelrichtlinien diesem Querschnittsanspruch gerecht werden, wird die zukünftige Praxis weisen. Wünschenswert wäre in jedem Fall die Erweiterung des Projekts durch besondere Sammelschwerpunkte. So wäre der Versuch denkbar, alle einen Ort betreffenden Websites von der Gemeindeverwaltung über Vereine, Firmen, Kirchengemeinden bis hin zu Privatanbietern zu sammeln oder aber bestimmte thematische Schwerpunkte (z.B. das Weltkulturerbe Mittelrheintal) anzugehen. Die Sammelrealität wird vermutlich auf einen Kompromiss zwischen einer unter qualitativen Aspekten betriebenen Auswahl und dem beschriebenen dokumentarischen Querschnittscharakter hinauslaufen.

Zur Zeit erfolgt die Auswahl der Websites zum einen durch Auswertung regionaler Linksammlungen, zum anderen werden bei der ohnehin erfolgenden Auswertung von regionalen Zeitungen für das Ermitteln der gedruckten Pflichtexemplare eine große Menge an URLs von Vereinen, Gesellschaften, Firmen, Gruppen und sonstigen Institutionen ermittelt.

¹³ Vgl. dazu die Projekte „Kompetenznetzwerk Langzeitarchivierung“, <http://www.ddb.de/professionell/projekte.htm#kompetenz> und „Langzeitarchivierung digitaler Dokumente in Deutschland: Initialzündung für die Erstellung eines gesamtdeutschen Konzepts“, <http://www.ddb.de/professionell/projekte.htm#lza>

4. Aussicht

Die eingesammelten Metadaten sollen vom zur Zeit beim HBZ im Aufbau befindlichen OAI Service Provider eingesammelt werden und diese Daten auch über die Digitale Bibliothek NRW recherchierbar sein.

Die Übernahme der Metadaten in den HBZ-Verbundkatalog ist ebenfalls möglich. Ein Skript, welches die Daten in das Verbundsystem übernehmen könnte, liegt bereits vor. Eine gemeinsame Suchoberfläche für die RPB und die landeskundlichen Metadaten soll mittelfristig geschaffen werden. Diese Konstruktion ist zudem unproblematisch, da die Rheinland-Pfälzische Bibliographie nicht mehr in gedruckter Form erscheint.

Geplant ist zur Zeit, die bereits erfassten Websites im Abstand von sechs Monaten erneut einzusammeln. Ob sich dieser Sammelrhythmus in der Praxis bewährt und aufrechtzuerhalten ist, hängt von den Speicherkapazitäten sowie dem Dokumentationszweck ab. So ist durchaus klar, dass sich beispielsweise das Webangebot einer Partei während eines Wahlkampfes bei einem solchen Turnus nicht befriedigend dokumentieren läßt.

Verabredet ist eine Kooperation mit der DDB, in der die Regionalbibliotheken sich um die Sammlung der Netzpublikationen außerhalb des Buchhandels, die DDB um die im Buchhandel erschienenen bemüht. Gegenseitig stellen sich die Partner anschließend die Dokumente sowie die zugehörigen Metadaten zur Verfügung. Inzwischen hat sich eine Arbeitsgruppe mit Vertretern aus DDB, HBZ, Bibliotheksservice-Zentrum Baden-Württemberg und den Landesbibliotheken in Karlsruhe, Stuttgart und Koblenz konstituiert, die gemeinsame Metadatenvorgaben und technische Austauschformate vereinbaren soll.

Eine verstärkte Kooperation der Regionalbibliotheken auf diesem Feld wäre daher wünschenswert: insbesondere die Landesbibliotheken aus NRW sollten nicht außen vor bleiben.