Langzeitarchivierung von Online-Publikationen an Regionalbibliotheken: Das Projekt ‚Baden-Württembergisches Online-Archiv' (BOA)

Heidrun Wiesenmüller

Elektronisches Publizieren im Internet ist heute in vielen Bereichen Realität geworden – bei der ‚grauen Literatur' vielleicht mehr noch als bei den Verlagen. Immer häufiger erhält man z. B. bei Zeitschriften die Auskunft „erscheint nur noch online". Dabei sind es keineswegs nur Mitteilungsblätter und Jahresberichte kleiner Vereine, die nicht mehr gedruckt werden – auch eine Institution wie beispielsweise das Statistische Landesamt Baden-Württemberg will wichtige Publikationen wie die ‚Statistischen Berichte' künftig ausschließlich im Internet anbieten. Ebenso ist es bei der Arbeit in den bibliographischen Stellen (z. B. Landesbibliographie, Hölderlin-Bibliographie), Sonderabteilungen und Fachreferaten alltäglich geworden, auf einschlägige und inhaltlich substanzielle Online-Ressourcen zu stoßen, die man gerne festhalten und dauerhaft archivieren würde. Der praktische Umgang damit war bisher jedoch meist von einer gewissen Hilflosigkeit geprägt; die Materialien wurden – mit wenigen Ausnahmen[1] – allenfalls in Linklisten aufgenommen oder gar auf Papier ausgedruckt.

Gleichzeitig wuchs in den vergangenen Jahren die Einsicht, dass – trotz aller technischen Schwierigkeiten und der allgegenwärtigen Personalknappheit – der Sammel- und Archivierungsauftrag der regionalen Pflichtexemplarbibliotheken vor den Netzpublikationen nicht Halt machen dürfe. Die beiden baden-württembergischen Landesbibliotheken in Karlsruhe und Stuttgart wollten sich der neuen Verantwortung offensiv stellen, wobei sich als natürlicher Kooperationspartner das Bibliotheksservice-Zentrum in Konstanz anbot. Im Oktober 2002 – kurz vor dem von Der Deutschen Bibliothek initiierten Workshop ‚Langzeitverfügbarkeit elektronischer Dokumente' – wurde deshalb eine Vereinbarung zwischen der Badischen Landesbibliothek (BLB), der Württembergischen Landesbibliothek (WLB) und dem Bibliotheksservice-Zentrum Baden-Württemberg (BSZ) geschlossen. Das Ziel war die gemeinsame Entwicklung einer technischen Plattform und eines Geschäftsgangs für die Erschließung und Speicherung relevanter Netzpublikationen. Seit Herbst 2003 trägt das Projekt den Namen ‚Baden-Württembergisches Online-Archiv' (BOA) und ist unter der URL

1 Zu nennen sind vor allem elektronische Hochschulschriften und das Angebot der ‚Elektronischen Zeitschriftenbibliothek',
URL: <http://rzblx1.uni-regensburg.de/ezeit/ezb.phtml>.

<http://www.boa-bw.de> erreichbar. Die technische Entwicklung wird vom Ministerium für Wissenschaft, Forschung und Kunst des Landes Baden-Württemberg mit einer Anschubfinanzierung gefördert; für die beteiligten Bibliotheken gibt es allerdings derzeit weder Sondermittel noch zusätzliches Personal.

Grundprinzipien des BOA-Projekts

Im Zentrum des Interesses stehen Netzpublikationen, die den Charakter eines Pflichtexemplars besitzen oder von landeskundlicher Bedeutung bzw. für die Sondersammlungen der beteiligten Bibliotheken relevant sind. Zunächst soll sich die Sammelaktivität dabei nicht auf kommerzielle Angebote, sondern auf frei verfügbare Internet-Ressourcen konzentrieren. Die Methoden, solche Objekte aufzuspüren, sind vielfältig: Man findet sie z. B. über Presseberichte und Linksammlungen, aber auch sehr häufig bei ohnehin anfallenden Recherchen in Abteilungen wie der Landesbibliographie oder der Pflichtstelle. Detaillierte Sammelrichtlinien müssen freilich erst erarbeitet werden. Deshalb ist es in der Anfangsphase nötig, in Frage kommende Objekte genau zu prüfen. Der im BOA-Projekt verfolgte Ansatz stellt dabei Qualität über Quantität. Im Bereich der Websites beispielsweise geht es weniger um Homepages von Städten, Vereinen o.ä., sondern eher um umfassende thematische Angebote[2].

Von Anfang an war klar, dass die archivierten Dokumente einen integralen Teil des Bibliotheksbestandes bilden sollten. Daraus ergibt sich zum einen, dass sie – genau wie die konventionellen Materialien – eine vollwertige Formal- und Sacherschließung erhalten müssen. Zum anderen sollten sie nicht nur in einer getrennten Datenbank, sondern auch über die lokalen OPACs recherchierbar sein. Bei der Erfassung darf es keine Doppelarbeit geben; nötig ist vielmehr ein reibungsfreier und möglichst ‚schlanker' Geschäftsgang. In Baden-Württemberg entschloss man sich deshalb (anders als im rheinland-pfälzischen Projekt eDOWEB[3]) dazu, auch die Online-Publikationen in den zentralen Nachweisinstrumenten zu katalogisieren – also im Südwestverbund (SWB) bzw. in der Zeitschriftendatenbank (ZDB). Von dort werden die Titelaufnahmen in das Depotsystem übernommen. Auch etwaige spätere Änderungen werden nur im Verbund bzw. der ZDB vorgenommen; die Aktualisierung im BOA-System erfolgt automatisch.

2 Z. B. <http://www.historisches-wuerttemberg.de> mit Informationen zu Burgen, Schlössern, Klöstern etc. oder <http://www.literaturland-bw.de> über literarische Museen, Archive und Gedenkstätten.

3 URL: <http://www.rlb.de/edoweb.html>. Vgl. Jendral, Lars et al.: Archivierung von landeskundlichen Netzpublikationen: ein Projekt der Rheinischen Landesbibliothek und des Hochschulbibliothekszentrum Köln, in: Prolibris 2003, Heft 4, S. 199-203. Auch im BIBLIOTHEKSDIENST erscheint demnächst ein Artikel dazu.

Bisheriger Projektablauf

Die konzeptionellen Details wurden in einer Arbeitsgruppe erarbeitet, in der neben den drei Projektpartnern weitere Landesbibliotheken aus der Verbundregion sowie die Landesarchivdirektion Baden-Württemberg vertreten sind – denn von Anfang an war das Projekt auf Kooperation und Nachnutzung angelegt. Für den ersten, im April 2003 fertiggestellten Prototyp griff das BSZ auf die bewährte OPUS-Software zurück. Allerdings ist OPUS, das bisher vor allem im Bereich elektronischer Hochschulschriften zum Einsatz kommt, nicht für die Verwaltung hierarchisch verknüpfter Elemente ausgelegt. Für das BOA-Projekt war dies eher ungünstig, da man im Internet sehr häufig auf Zeitschriftenartiges stößt. Hierunter fallen nicht nur klassische, in Einzelheften erscheinende Zeitschriften, sondern auch Websites.

Abb. 1: Detailansicht einer BOA-Aufnahme mit Link zum Dokument

Eine Lösung des Problems versprach die Java-Anwendung ESEM, die vom BSZ für die Organisation elektronischer Semesterapparate an der UB Konstanz[4] entwickelt worden war. Untergeordnete Objekte lassen sich damit sehr komfortabel in beliebig vielen Hierarchieebenen anlegen und verwalten. Nach einer entsprechenden Funktionserweiterung präsentierte das BSZ den Betei-

4 URL: <http://esem.bsz-bw.de/Web/index.jsp>.

ligten im September 2003 einen neuen Prototyp auf der veränderten Software-Basis; als Datenbank-System kommt ORACLE zur Anwendung. Die Entwicklungs- und Programmierarbeiten sind zwar noch nicht abgeschlossen, aber doch schon weit fortgeschritten. Seit Januar 2004 werden Aufnahmen in der Echt-Version von BOA angelegt (Abb. 1). Zugleich ist eine Test-Installation in Betrieb, auf der neue Funktionalitäten eingespielt und getestet werden können, ehe sie in die Echt-Version übernommen werden. Bei einem Treffen von Vertretern der regionalen Pflichtexemplarbibliotheken am 21. Januar 2004 in Fulda wurde das System erstmals einer größeren Fachöffentlichkeit vorgestellt.

Die Erfassung im BOA-System

Wie sieht die Erfassung einer Online-Publikation nun in der Praxis aus? Ist die Entscheidung für eine Archivierung gefallen, so wird das Dokument zunächst im Verbundkatalog bzw. in der ZDB nach RAK-NBM katalogisiert. Aus dem Katalogisat wird (über vom SWB bzw. PICA/Iltis vorgehaltene Routinen) eine Download-Datei im MAB2-Format generiert. Um diese in das BOA-System einzulesen, muss der Bearbeiter – nachdem er sich authentifiziert hat – lediglich Dateinamen und Pfad in einem Web-Formular angeben (Abb. 2). Auch für alle weiteren Bearbeitungsschritte genügt ein Browser; es muss keine spezielle Software installiert werden.

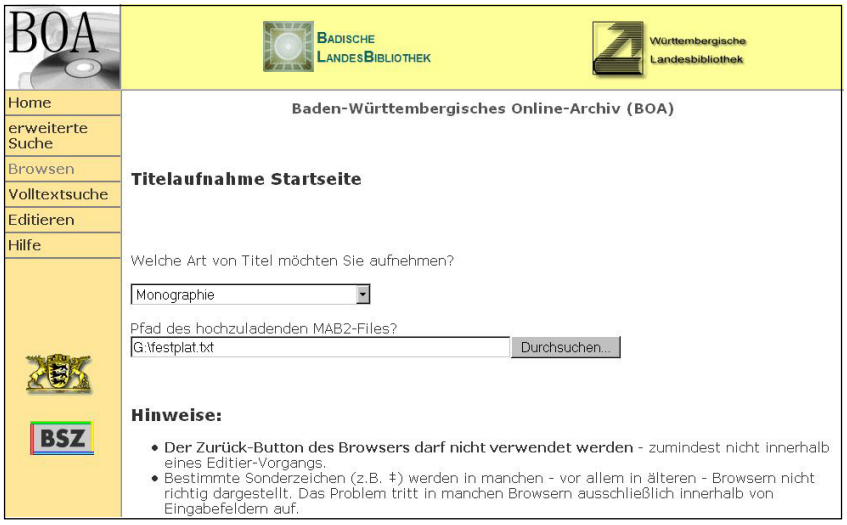

Abb. 2: Einlesen einer MAB-Datei in BOA

Anfangs war geplant, die bibliographischen Daten innerhalb des BOA-Systems im Dublin-Core-Format zu halten, doch wurde diese Idee wieder verworfen. Stattdessen liegen sie nun auch intern in der MAB-Struktur vor. Als mögliches Ausgabeformat soll Dublin Core natürlich trotzdem unterstützt werden.

Gemäß dem Verbundstandard werden die erfassten Objekte nach RSWK verschlagwortet, wobei auch die Schlagwortketten aus dem SWB nach BOA übernommen werden. Zusätzlich werden DDC-Sachgruppen Der Deutschen Bibliothek sowie Notationen der Landesbibliographie von Baden-Württemberg vergeben (derzeit im BOA-System selbst, künftig evtl. im Verbund), die ein thematisches Browsing ermöglichen. Bei Bedarf können weitere Systematiken (z. B. die Regensburger Verbundklassifikation) integriert werden. Die bibliographische Beschreibung wird im BOA-System durch verschiedene Verwaltungsdaten ergänzt: Hierunter fällt etwa die Angabe eines Termins, an dem eine einmal erfasste Website erneut abgespeichert werden soll. Das regelmäßige Einsammeln solcher Objekte soll künftig weitgehend automatisiert erfolgen.

Ebenfalls integriert ist eine Art ‚elektronischer Laufzettel' mit bestimmten Geschäftsgangsinformationen. Besonders wichtig ist dabei das Vorliegen der Genehmigung des Anbieters einer Ressource, ohne die die Speicherung in BOA bei der geltenden Rechtslage nicht zulässig ist. Das Einholen dieser Genehmigung stellt natürlich einen Zusatzaufwand dar. Umso mehr ist zu hoffen, dass die Pflichtexemplargesetze der Länder möglichst bald auch Regelungen für Netzpublikationen beinhalten werden. Ein entsprechender Musterentwurf wurde 2003 von der AG Regionalbibliotheken bei der Kultusministerkonferenz eingebracht.

Abspeichern von Online-Publikationen

Das Abspeichern der Netzpublikation selbst geschieht auf Knopfdruck: Es muss nur die zugehörige URL in ein Feld eingetragen werden, um die Objekte auf den Archivserver im BSZ zu übertragen. Neben PDF-Dokumenten können auch HTML-Objekte archiviert werden. Dabei handelt es sich entweder um komplette Websites oder um Teile davon, welche wiederum eingebettete Objekte in anderen Formaten (z. B. Graphiken) enthalten können.

Im rheinland-pfälzischen eDOWEB-Projekt werden Objekte im HTML-Format in einem getrennten System zur Verfügung gestellt und in der Regel weniger tief erschlossen als solche im PDF-Format. BOA macht diesen Unterschied nicht: Zwar wirken PDF-Ressourcen auf den ersten Blick oft solider und ‚habhafter', weil sie den aus der Print-Welt bekannten Publikationstypen am ehesten entsprechen. Doch auch im HTML-Format kommen eindeutige Publi-

kationen wie Reiseführer, Lexika oder Ausstellungskataloge[5] vor. Bei Online-Zeitschriften stößt man sogar nicht selten auf die Situation, dass ein Teil der Hefte als PDF, der andere im HTML-Format angeboten wird. Das Format wird daher im BOA-Projekt nicht als primäres Unterscheidungskriterium betrachtet.

Das Abspeichern von HTML-Ressourcen erfolgt mit dem Offline-Browser HTTrack, der auch bei eDOWEB sowie bei Projekten Der Deutschen Bibliothek zum Einsatz kommt. HTTrack ‚grast' sich sozusagen Link für Link durch das zu archivierende Objekt und kopiert jede Seite, wobei die interne Struktur der Website erhalten bleibt – das Ergebnis ist eine genaue Spiegelung des Originalangebots. Im BOA-System lässt sich einstellen, wieviele Ebenen eingesammelt werden, ob auch externe Links verfolgt werden sollen, welche maximale Größe die Dateien haben dürfen etc. (Abb. 3).

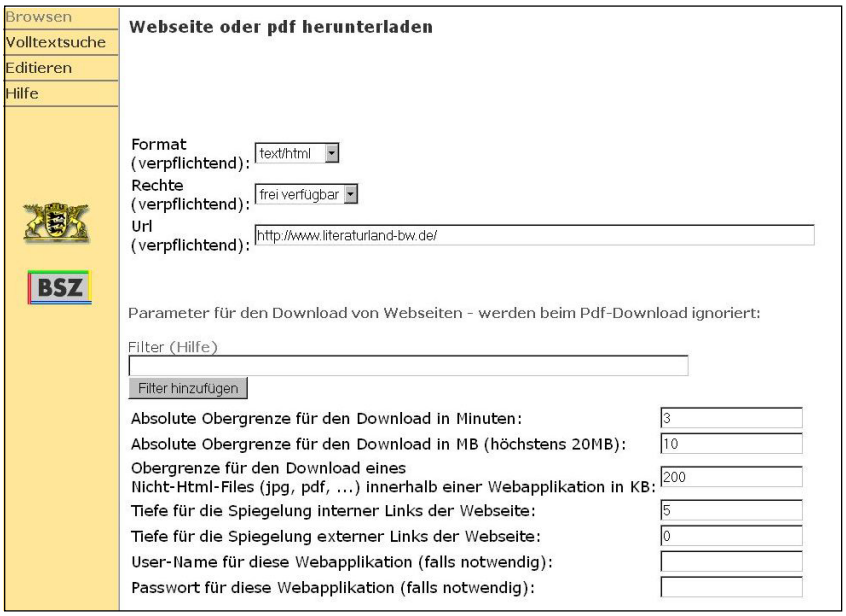

Abb. 3: Einstellen von Parametern für HTTrack

5 z. B. <http://www.s-line.de/homepages/ebener/index.htm> (Kleines Lexikon zur Geschichte in Baden und Württemberg, philatelistisch unterstützt), <http://www.burgen.strasse-online.de/index.html> (Burgenstraße zwischen Mannheim, Heidelberg, Rothenburg o. d. Tauber, Nürnberg, Bayreuth und Prag: ein Internet-Reiseführer), <http://www.karlsruhe.de/Kultur/MLO/katalog/index.htm> (Geschichte der Literatur am Oberrhein: ein Querschnitt; Katalog zur ständigen Ausstellung des Museums für Literatur am Oberrhein).

Nach dem Abspeichern kann man die Archivkopie begutachten und bei Bedarf den Vorgang mit geänderten Parametern wiederholen. In der überwiegenden Zahl der Fälle ist das Ergebnis gelungen, doch kommt es gelegentlich auch zu Fehlern. Ob evtl. eine andere Software besser geeignet ist als HTTrack, wird deshalb noch zu prüfen sein. Generell nicht möglich ist derzeit die Übernahme ganzer Web-Datenbanken. Auch dynamisch generierte Websites bereiten Schwierigkeiten, die voraussichtlich erst in einer späteren Projektstufe gelöst werden können.

Soll ein neues Heft einer Zeitschrift abgespeichert werden, so wird zunächst in BOA ein entsprechendes untergeordnetes Objekt angelegt (Abb. 4) und danach die zugehörige Netzpublikation abgespeichert – ein Vorgang, der kaum länger dauert als das Eintragen in einen konventionellen Kardex. Auch die Bestimmung der URN (Uniform Resource Name) als dauerhafte Adresse übernimmt das BOA-System; diese muss dann noch im Verbund bzw. in der ZDB nachgetragen und bei Der Deutschen Bibliothek registriert werden.

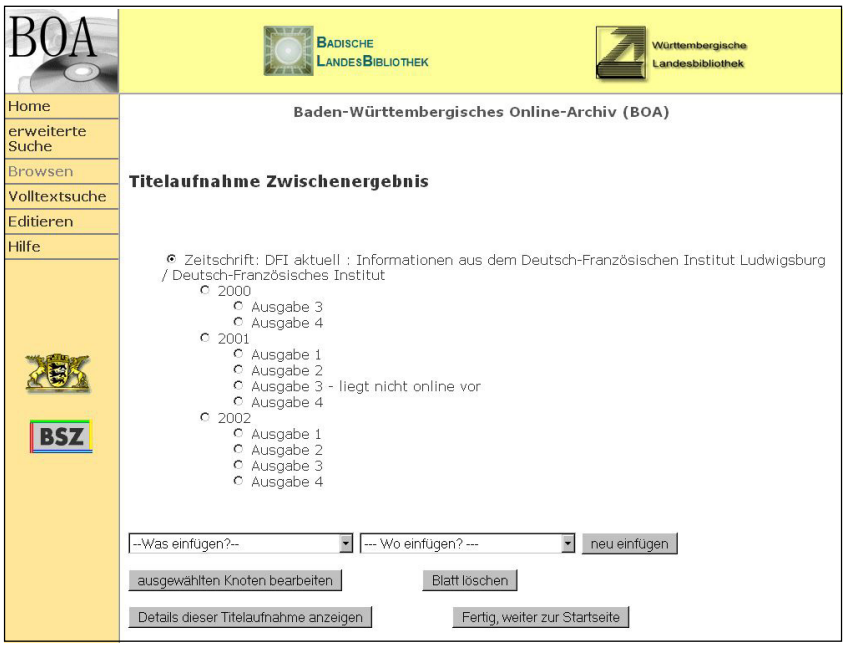

Abb. 4: Anlegen eines untergeordneten Objekts (z. B. Zeitschriftenheft)

Die Benutzersicht

Sobald bei den erfassten Dokumenten eine ‚kritische Masse' erreicht und die Benutzersicht optimiert ist, soll BOA auch der breiten Öffentlichkeit bekannt gemacht werden. Dabei wird es grundsätzlich zwei Möglichkeiten geben, auf die archivierten Online-Publikationen zuzugreifen: Zum einen über eine ganz ‚normale' Recherche im lokalen OPAC oder im Südwestverbund bzw. der ZDB – von der Trefferanzeige führt dort ein Link zum zugehörigen BOA-Objekt. In diesem Fall erfüllt sich also der Wunschtraum vieler Benutzer – im Katalog nicht nur den Literaturnachweis, sondern auch das Dokument selbst zu finden! Zum anderen kann natürlich auch direkt im BOA-System recherchiert werden. Zusätzlich zu einer OPAC-ähnlichen Suchmaske gibt es hier die Möglichkeit, nach den Notationen der integrierten Systematiken thematisch zu browsen. Darüber hinaus lassen sich die archivierten Objekte nach Materialart getrennt durchsuchen. Über eine Suchmaschine wird zudem auch eine Volltextsuche in den Dokumenten möglich sein.

Zwar kann man sich in der Archivkopie eines HTML-Objekts genauso bewegen wie im Original, doch gibt es notwendigerweise Grenzen – nämlich dann, wenn man einen Link anklickt, der nicht mitgespeichert wurde. Sofern die Seite noch existiert, verlässt der Benutzer in einem solchen Fall das BOA-Objekt und wechselt ins freie Internet.

Langzeitarchivierung

Mit dem Abspeichern einer Netzpublikation in BOA ist freilich nur der erste Schritt getan: Gesichert ist damit zwar die aktuelle Benutzung der Ressource, auch wenn der Anbieter sie wieder vom Netz nimmt oder verlagert, aber noch nicht die Langzeitverfügbarkeit. Diese kann nur über Methoden wie Emulation und Migration erreicht werden, wofür entsprechende Systeme erst aufgebaut werden müssen. Unabdingbar ist es jedoch, schon bei der Archivierung nicht nur bibliographische, sondern auch bestimmte technische Metadaten zu erfassen. Diese bilden die Grundlage für zukünftige Datenpflege-Aktivitäten zur Langzeitarchivierung.

Zu diesen technischen Metadaten gehören u. a. Dateigrößen und Formate, aber auch Informationen darüber, in welchem Programm das Objekt darstellbar ist und sogar – wenn möglich – mit welcher Software es erstellt wurde. Ist beispielsweise aus den Metadaten abzulesen, dass eine Ressource zum Zeitpunkt ihrer Archivierung in Mozilla 1.4 darstellbar war, so würde auch in einer fernen Zukunft eine diesem Browser entsprechende Systemumgebung emuliert werden. Für die Migrationsmethode ist hingegen die Kenntnis des Erstellungsprogramms wichtig – denn die Daten müssen rechtzeitig aktualisiert werden, ehe dieses obsolet wird. Es versteht sich, dass ein möglichst großer

Teil der technischen Metadaten maschinell ausgelesen oder als Defaultwert vorgegeben werden muss, um den Aufwand für die Bearbeiter in den Bibliotheken gering zu halten.

Seit Oktober 2003 plant eine Arbeitsgruppe aus Der Deutschen Bibliothek und Vertretern der Projekte eDOWEB (Rheinische Landesbibliothek, Hochschulbibliothekszentrum) und BOA (BLB, WLB, BSZ) die technischen Details für einen Datenaustausch im Rahmen einer Kooperation bei der Langzeitarchivierung elektronischer Pflichtexemplare. In diesem Zusammenhang werden die Erfordernisse technischer Metadaten für die Langzeitarchivierung ausführlich diskutiert; ein Entwurf liegt vor.

Ausblick

In den eineinhalb Jahren, die seit dem ‚Startschuss' für BOA vergangen sind, ist von den Projektpartnern bereits beträchtliche und sehr erfolgreiche Arbeit geleistet worden: Eine einsatzfähige technische Plattform liegt vor, und auch die reibungsfreie Einbettung in überregionale Entwicklungen ist durch den regelmäßigen Austausch mit Der Deutschen Bibliothek und eDOWEB gelungen.

Dennoch bleibt noch vieles zu tun: Die Technik muss – gerade mit Blick auf die Langzeitverfügbarkeit – optimiert und weiterentwickelt werden. Um die rechtliche Grundlage zu verbessern, ist die Revision der Pflichtexemplargesetze voranzutreiben. Es müssen Erfahrungen mit dem neuartigen Bibliotheksgut gesammelt und Richtlinien für die Auswahl entwickelt werden. Und nicht zuletzt muss in den Bibliotheken Arbeitskraft und Zeit dafür aufgebracht werden, relevante Netzpublikationen in größerer Zahl zu identifizieren und in BOA zu archivieren. Da der Aufwand für die konventionellen Medien nicht geringer werden wird und gleichzeitig die Personalressourcen weiter schrumpfen, stellt dies eine gewaltige Herausforderung dar.