

## Dauerhafte Verknüpfung von Bibliothekskatalogen mit Internetressourcen

### Vorarbeit für ein Konzept zur Aktualisierung von URLs oder Ersetzung durch URNs

**Henning Klauß**

Der Aufsatz gilt der Frage, wie die Verknüpfung von einem Bibliothekskatalog zu Internetressourcen stabil gehalten werden kann. Die Behebung dieses Problems wird bisher punktuell in einigen Bibliotheken oder auch Verbänden bearbeitet, ist aber bisher nicht allgemein gelungen. Für das dazu zu entwickelnde überregionale Konzept sollen einige lokale Erfahrungen dargestellt werden. Intendiert ist ein (zunächst) deutschlandweites Konzept.

Vergegenwärtigen wir uns zunächst allgemeine, d.h. nicht private, sondern gesellschaftliche Bestimmungen der Lebenswelt des Benutzers: Wir leben in einer Zeit, die u.a. durch die sogenannte Informationsflut und durch Zeitökonomie (die oftmals als Hast in Erscheinung tritt) gekennzeichnet ist. Das bedeutet für den Benutzer: Er will alles und zwar sofort!

Dem Wunsch, „alles“ zu bekommen, kann man im strengen Sinne nur gerecht werden, wenn man dem Benutzer auch alles gibt – dann aber ist er prompt verloren. Infolgedessen ist es auch hinsichtlich der Zeitökonomie notwendig, sinnvolle Strategien einer gezielten Informationssuche anzubieten, bei der ggf. auch viel wegggeschnitten wird, um das Wesentliche in den Blick zu bekommen – und das ist es ja auch, was ein Benutzer meint, wenn er „alles“ sagt, denn er will ja nicht „irgendwie alles“, sondern alles zu einem bestimmten Thema oder von einer bestimmten Person oder Körperschaft. Eine „gezielte Informationssuche“ erfordert zunächst, an möglichst viele Informationen anknüpfen zu können; das betrifft z.B. die Fragen des Katalogregelwerkes (ob darin z.B. auch die Verzeichnung des zweiten Herausgebernemens vorgesehen ist), ob Aufsätze in Sammelbänden und Zeitschriften verzeichnet werden usw., so dass mit diesem Namen oder diesen Stichwörtern gesucht werden kann. Darüber hinaus ist von Belang, ob die Titelaufnahmen in anderen Feldern (Schlagwörter, Fußnoten, Abstracts...) hinreichend üppig gefüllt sind, um sie in der feldübergreifenden freien Suche im ersten Retrievalschritt ermitteln zu können. Erst wenn die Bedingungen für die gehaltvolle Ermittlung einer Treffermenge erfüllt sind, kann im nächsten Schritt eine sinnvolle Reduktion durchgeführt werden.

Das Ergebnis einer Katalogrecherche besteht zunächst aus Titelaufnahmen. Diese sollten so beschaffen sein, dass der Benutzer am Bildschirm schon viel erledigen oder zumindest vorentscheiden kann. Natürlich können wir es nicht komplett ver-

meiden, dem Benutzer eine Titelaufnahme zu liefern, die ihn zu unrecht motiviert, das damit beschriebene Buch am Regal aufzusuchen und dann davon enttäuscht zu sein. Aber wenn es möglich ist, den Informationsgehalt am Bildschirm so zu gestalten, dass Frusterlebnisse seltener, Erfolgserlebnisse häufiger werden, sollten wir die Möglichkeit dazu nutzen.

Eine Möglichkeit, dem Benutzer unnötige Wege zu ersparen, ist dann gegeben, wenn zu Medien Internetressourcen vorliegen und eine geeignete Verknüpfung es dem Benutzer erspart, dafür die Recherche ein weiteres Mal in einer Suchmaschine abzusetzen.

Hierfür gibt es zunächst zwei Möglichkeiten: Entweder der derzeitige Inhalt einer Internetquelle wird in ein Feld der Titelaufnahme kopiert. Das würde den Katalog enorm aufschwemmen (insbesondere im Fall der Dissertationen) und ihn, sofern das Feld indexiert wird, langsamer machen. Zudem würde sich in gewisser Hinsicht ein Aktualisierungsproblem in extremer Schärfe ergeben, z.B. dann, wenn die aktuellen Inhaltsverzeichnisse von Zeitschriften protokolliert werden. Denn die ändern sich im Gegensatz zu „fertigen“ Texten wie Dissertationen ständig. Vorteil: Der Benutzer könnte den Katalog ggf. (ohne Einsatz von weiteren Suchmaschinen-Technologien<sup>1</sup>) für eine Volltextsuche nutzen. Die andere Möglichkeit der Verknüpfung besteht darin, nicht den derzeitigen Inhalt einer Internetseite selbst zu importieren, sondern „nur“ den Verweis darauf, so dass der Benutzer keine statischen, sondern wechselnde, i.d.R. aktuelle Informationen bekommt. Letzteres hat sich durchgesetzt, indem z.B. URLs protokolliert werden.

Seit einigen Jahren (in der UB der Europa-Universität Viadrina seit 1998) ist es üblich, die Verknüpfung von Bibliothekskatalog und Internetressourcen durch Angabe der URLs zu bewerkstelligen. Hierfür sind folgende Fragen zu klären:

1. In welchen Fällen wird eine URL protokolliert? Die Internetressource sollte in einem möglichst konkreten Zusammenhang zu einer im Katalog vorhandenen Titelaufnahme stehen, also der Volltext in der gleichen oder einer anderen Sprache, titelspezifische Meta-Information, wie z.B. Abstract, Rezension, Inhaltsverzeichnis, Verlagsbeschreibungen ...; Internetressourcen, die nur „irgendwie“ zum Thema passen, haben an dieser Stelle m.E. nichts zu suchen, denn es könnte für manche Benutzer der irreführende Eindruck entstehen, an dieser Stelle des OPACs würde eine vollständige Literaturliste zu einem Thema vorliegen. Hierzu gehört auch die Frage, ob bei Schriftenreihen die URL im Gesamttitel oder bei den Stücktiteln verzeichnet wird; ggf. an beiden Stellen? Zudem: Werden auch personen- und körperschaftsspezifische URLs protokolliert?

---

1 Der Einsatz solcher Technologien ermöglicht es, den Inhalt der WWW-Seiten, auf die in den Titelaufnahmen verwiesen wird, in den Index einzubinden, ihn also recherchierbar zu machen.

2. Wer gibt die Anregung für eine zu protokollierende URL: AK, Fachreferenten, Benutzer...? Sind für die abschließende Entscheidung der Aufnahme einer URL die Fachreferenten zuständig? (Achtung: Das könnte das Durchlaufen des Geschäftsgangs sehr verzögern!)
3. In welchem Fall wird welches Feld des Bibliothekssystems genutzt? Hinweise auf Parallelveröffentlichungen kommen eventuell in ein anderes Feld als Hinweise auf Abstracts, Rezensionen, Titelblätter, Inhaltsverzeichnisse ...
4. Formulierung des Textvorspannes bzw. „Einbettung“ von URLs: „Rezension unter <http://...>“, „Abstracts unter <http://...>“, „U.d. URL <http://...> sind einzelne Aufsätze in dieser Schriftenreihe (ab Band 500) recherchierbar“ oder ganz einfach, ziemlich informationsfrei und daher kaum zu verändern, „s.a. <http://...>“
5. Wie werden diese URLs aktualisiert? Eine Aktualisierung kann sowohl Änderung als auch Löschung bedeuten, Letzteres kann z.B. irgendwann bei Links auf Verlagsbeschreibungen erforderlich sein: Wenn der Verlag ein bestimmtes Buch nicht mehr verkauft, wird er die Beschreibung aus dem Netz nehmen.

Die Frage 2. wird i.d.R. bibliotheksintern, die Fragen 1., 3. und 4. werden, soweit das gegeben ist, i.d.R. seitens der Bibliotheksverbände geklärt. Nur die letzte der hier aufgeführten Fragen soll im Folgenden betrachtet werden, denn die Aktualisierung verdankt sich bisher weitgehend individuellen Leistungen und Zufälligkeiten. Dieser Zustand sollte verbessert werden.

Die wohl „hässlichste“ Eigenschaft von URLs ist deren Kurzlebigkeit, sprich die zeitliche Begrenztheit ihrer Gültigkeit. Da wird ein Server umbenannt, die Verzeichnisstruktur geändert, eine einzelne Web-Seite nicht mehr gepflegt oder einfach gelöscht; immer wieder heißt es, eine Seite sei „umgezogen“ [obwohl doch nur Subjekte, Menschen umziehen, während Objekte wie z.B. WWW-Seiten „umgezogen werden“]... . Auch Kleinigkeiten wie z.B. die Veränderung der Extension im Dateinamen von „\*.htm“ zu „\*.html“ oder umgekehrt wirken sich desaströs aus.

Wenn sich in einem Bibliothekskatalog tote Links häufen, nehmen Benutzer diese Elemente der Titelaufnahmen eventuell irgendwann pauschal nicht mehr als Bereicherung wahr: „Schon wieder veraltete WWW-Müll-Informationen!“ Weil die URLs aber Bereicherungen sind bzw. sein könnten, ist deren beständige Aktualisierung sicherzustellen. Wenn dies nicht gewährleistet werden kann, kann die Menge toter Links und die Reaktion der Benutzer auf diesen Umstand Bibliotheken veranlassen, die URLs enthaltenden Felder aus der Anzeige im OPAC herauszunehmen. Die Verknüpfung von Bibliothekskatalog und Internetressourcen ist damit weitestgehend hinfällig.

Die Aktualisierung der Titelaufnahmen hinsichtlich der darin enthaltenen URLs wird in unserer Bibliothek bewirkt, indem die im Katalog vorfindlichen URLs extrahiert, in eine Datei zusammengefasst und von einem Linkchecker überprüft werden. Wir verwenden hierfür „GNU Wget“ (<http://www.gnu.org/software/wget>)

/wget.html), eine Open-source-Software. Die durch diese Prüfung entstehenden Seiten werden in unserer UB auf die kurze Form gebracht:

<http://www.elsevier-international.com/catalogue/title.cfm?ISBN=0125649150>  
failed: Connection timed out. Giving up.

<http://jiv.sagepub.com/> failed: Connection timed out. Giving up.

<http://vsd.pennnet.com/home.cfm> connected. Giving up.

<http://w210.ub.uni-tuebingen.de/dbt/intro> connected. 404 Not Found

<http://w210.ub.uni-tuebingen.de/dbt/volltexte/> connected. 403 Forbidden

<http://www-docs.tu-cottbus.de/bibliothek/public/katalog/>

connected. 403 Forbidden

<http://www.sourceoecd.org/> connected. 501 Illegal/Unknown Web Request

Bei diesen Kommentierungen ist allerdings zu berücksichtigen, dass die Meldung „forbidden“ nicht Ausdruck einer unter allen Bedingungen unzugänglichen WWW-Seite ist: Es kann sich evtl. auch darum handeln, dass nur seitens bestimmter Bibliotheken die notwendigen Lizenzbedingungen vorliegen (IP-Check) oder aber der angesprochene Server Anfragen von Linkcheckern (falls die sich, wie Wget, korrekt „outen“) ablehnt. Die Kommentare sind also nicht einfach hinzunehmen, sondern interpretationsbedürftig.

Diese Überprüfung haben wir einige Jahre nicht nur bei Titelaufnahmen, sondern auch bei den Personen- und Körperschaftendateien gemacht. Da sich die Überprüfung dieser beiden Dateien als wenig fruchtbar erwies, haben wir uns auf die Überprüfung der URLs in Titelaufnahmen beschränkt.

Da nicht alle URLs in indexierten Feldern stehen, sich in diesen Fällen daher nicht umstandslos ermitteln lässt, in welcher Titelaufnahme eine URL eingebunden ist, haben wir eine weitere WWW-Seite eingerichtet, der zu entnehmen ist, welche URL sich in welcher Titelaufnahme (Katalognummer) steht.

<http://tar.sagepub.com/> 4295538

<http://taylorandfrancis.metapress.com/> 4288496,4288571,4287737

<http://tcbh.oxfordjournals.org/> 4288783

<http://tcn.sagepub.com/> 4289802

<http://tcp.sagepub.com/> 4288089

<http://tcpt.alexanderstreet.com> 4228085

Wenn der Linkchecker eine URL als problematisch identifiziert hat, wurde die zugehörige Titelaufnahme entweder direkt im Bibliothekssystem oder mittels dieser WWW-Seite ermittelt. Je nachdem, ob die URL in einem exklusiv lokalen oder aber einem verbundrelevanten Feld stand, wurde die Korrektur im Lokal- oder im Verbundkatalog durchgeführt.

In leider nicht seltenen Fällen mussten kurz zuvor aktualisierte URLs ein weiteres Mal aktualisiert werden, da deren URL erneut geändert wurde.

Mit einem Linkchecker ist allerdings nur zu prüfen, ob die angegebenen URLs noch vorhanden sind; ob sich der Inhalt der unter der URL zur Verfügung gestellten Informationen ändert (wenn z.B. statt Volltexte nur noch Abstracts oder andere Erscheinungsjahre angeboten werden, Passwörter abgefragt werden ...), lässt sich so nicht automatisch überprüfen. Um bei weiterhin intakten URLs keinen künftig eventuell falschen Textvorspann zu formulieren, ist es daher ratsam, den Textvorspann hinreichend allgemein zu formulieren (also nicht z.B. „Inhaltsverzeichnis ab Band 300“, sondern besser „Inhaltsverzeichnis neuerer Bände“). Wenn sich ein Link nicht umstandslos aktualisieren, d.h. reparieren lässt, ist es die Aufgabe des jeweiligen Fachreferenten, hier Abhilfe zu schaffen. Praktisch sieht das so aus, dass im Fall von relevanten Links die aktuelle URL aufwendig ermittelt wird. Weniger relevante Links wurden nur aktualisiert, wenn diese sich halbwegs umstandslos ermitteln ließen, ansonsten wurden sie gelöscht.

Soweit zu dem, wie wir es bisher gemacht haben. Die Menge der zu überprüfenden Links ist aber in den letzten Jahren derartig gewachsen, dass trotz Komprimierung der zu prüfenden URLs<sup>2</sup> gravierende Belastungen für Mitarbeiter und Server entstanden sind und mit diesen aufwendig erzeugten Zwischenergebnissen nur noch stichprobenartig umgegangen wird. Ich weiß nicht, in wie vielen Lokal- bzw. Verbundsystemen eine solche Überprüfung durchgeführt und die Ergebnisse ggf. in die Verbunddatenbank eingetragen werden, aber es kann kein Dauerzustand sein, dass die diesbezügliche Qualität der Titelaufnahmen von dererlei Gegebenheiten abhängt, für die es eventuell lokal vereinbarte feste Zuständigkeiten gibt, die aber ansonsten, d.h. regional oder überregional eher den Eindruck von Zufälligkeiten erwecken!

Zukunftsfähiger als URLs sind Uniform Resource Names (URNs). Eine URN ist ein „Persistent Identifier“ [PI; Dauerhafte Bezeichner], der einen dauerhaften Zugriff und damit eine Zitierbarkeit auch bei einer eventuellen Änderung der URL gewährleistet. Die DNB hat im Rahmen des Projektes CARMEN [Arbeitspaket 4: Persistent Identifiers and Metadata Management in Science] im Projekt EPICUR ein URN-Management eingeführt, dessen Ziel es war, über einen Resolving-Mechanismus der URN die korrekte, d.h. aktuell gültige URL zuzuordnen.

Das Ganze basiert auf einer vorübergehenden Trennung von Identifizierung und Adressierung. Stabilität wird durch Vermittlung bewirkt: Durch den Katalog wird nicht der direkte Link zur Internetressource gelegt, sondern der zu einer bestimm-

---

2 Um z.B. nicht alle Working paper einer Körperschaft einzeln zu testen, haben wir die Datei so komprimiert, dass anstatt mehrerer hundert URLs nur eine einzige, die Einstiegs-URL getestet werden muss. – Allgemein: Wenn die alphabetische Auflistung der zu überprüfenden URLs serverspezifische Häufungen aufweisen, wird die Liste „eingedampft“, indem bestimmte URLs aus der Prüfung herausgenommen werden.

ten Stelle einer Vermittlungsagentur – und deren Aufgabe ist es, via Linkresolver die Verbindung zur aktuell gültigen URL herzustellen.

Im Rahmen ihres Zuständigkeitsbereiches ist die DNB verantwortlich für die Koordination des URN-Vergabeverfahrens, die Administration des URN-URL-Zuordnungsmechanismus sowie die Sicherstellung der Persistenz der URN.<sup>3</sup>

Das ist m.E. ein wichtiger Schritt, bedeutet auf dem derzeitigen Stand aber die Limitierung auf fest stehende Texte, z.B. Dissertationen. Die Integration von variablen WWW-Seiten (z.B. aktuelle Zeitschriften-Inhaltsverzeichnissen) ist derzeit mit URN noch nicht möglich<sup>4</sup>: Die Ersetzung von URLs durch URNs ist ein Gewinn an Sicherheit, aber jeder Gewinn an Sicherheit hat seinen Preis – und der besteht hier in der enormen Eingrenzung der Möglichkeit, den Bibliothekskatalog mit Internetressourcen zu verknüpfen.

URNs sind zukunftsfähiger als URLs, weil stabiler und damit arbeitssparend, sie sind bisher aber nicht hinreichend weit verbreitet, um schon im allein seelig machenden Sinne nutzbar zu sein: Es gibt derzeit viele relevante Internetressourcen, die entweder (noch) keine URN haben, obwohl sie nach dem derzeitigen Stand eine haben könnten oder aber deshalb, weil sie aufgrund der o.a. Umstände derzeit noch keine haben können.

Nett, wirklich nett wäre es natürlich, wenn im Zusammenhang von Änderungen oder Löschungen von URLs alle, die auf diese WWW-Seite einen Link gelegt haben, von den ändernden Personen ermittelt<sup>5</sup> und dann informiert werden würden. Was u.a. voraussetzt, dass erkenntlich ist, wer für eine bestimmte WWW-Seite verantwortlich ist. Aber das zu ermitteln, ist bei vielen WWW-Seiten sehr zeitaufwendig. Es ist daher nicht verwunderlich, dass diesbezüglich Meldungen so selten sind. Leider beflügelt die Nennung von Ansprechpartnern erfahrungsgemäß die diesbzgl. Kommunikationstätigkeit nicht sonderlich.

Abgesehen davon, dass ein solches Meldeverfahren ohnehin nicht wirklich gut klappt (ich selbst habe in den letzten 10 Jahren lediglich zwei Hinweise auf geänderte URLs erhalten): Es bedarf des Einsatzes von Technik und der Organisation von Arbeitszusammenhängen / Zuständigkeiten, um eine dauerhafte Aktualisierung von URLs zu bewerkstelligen; es bedarf einer Arbeitsweise, die nicht von Zufälligkeiten abhängt und unnütze Parallelarbeit verhindert.

---

3 Seit September 2001 werden URNs administriert. Seitdem wurden fast 3,7 Millionen URNs [Stand: Dez. 2010] registriert.

4 Es gibt allerdings Bestrebungen der DNB, für Zeitschriftentitel und Metadaten URNs zu vergeben. – Alle in OPUS verwalteten Dokumente haben eine URN; alle Metadaten zu Netzpublikationen, die die DNB ausliefert, verfügen über eine URN.

5 Die Syntax in z.B. *Google* oder *altavista* hierfür ist: „link:http://www...“

Das ist nur möglich, wenn, so weit es geht, auf URN umgestellt und der Rest der verbleibenden Arbeit auf Verbundebene geregelt wird.

Jede Bibliothek bzw. jeder Bibliotheksverbund muss sich also entscheiden, welchen Weg der Verknüpfung von Bibliothekskatalog und Internetressourcen sie beschreiten wird: URLs oder URNs. Aber als Tipp: Man kann das Eine tun, ohne das Andere zu lassen. Wenn möglich eine URN, wenn nicht, dann eben eine URL.

Abschließend folgender Verfahrensvorschlag: In Bibliotheksverbänden wird, soweit das nicht ohnehin der Fall ist, definiert, welche URLs für Bibliotheken sinnvoll sind. Von den protokollierten URLs werden die problematischen identifiziert. Diese wiederum werden maschinell (anhand von Unterfeldern) in zwei Gruppen aufgeteilt:

- a) weniger relevante, möglicherweise ohnehin nur kurzfristig intakte URLs (z.B. Verlagsbeschreibungen)
- b) dauerhaft relevante URLs.

Ich gehe davon aus, dass die zweite Gruppe die weitaus kleinere ist.

Die unter a) verzeichneten URLs werden nach einem festzulegenden Zeitraum erneut geprüft und dann aufgrund von Fehlermeldungen gelöscht. Wenn eine Bibliothek oder ein Bibliotheksverbund diese Prüfung ein weiteres Mal wiederholen möchte, kann das gerne getan werden. Damit endet der ausschließlich maschinell abzuwickelnde Anteil der Pflege.

Die unter b) verzeichneten URLs sind bedeutend arbeitsintensiver zu pflegen: Wenn die URL eine Verknüpfung von einer Zeitschrift mit einer Internetressource darstellt, wird eine Meldung an die ZDB gemacht. Wenn die URL eine Verknüpfung von der Titelaufnahme einer Monografie mit einer Internetressource darstellt, wird diese URL an die die WWW-Seite betreuende Institution mit der Bitte um URN-Vergabe gegeben; wenn das nicht möglich oder nicht gewollt ist, wird um die Nennung der aktuellen URL gebeten. Sobald das passiert und an den Bibliotheksverbund gemeldet worden ist, wird im Katalog die URL durch eine URN (notfalls die aktualisierte URL) ersetzt. Wenn sich binnen einer vereinbarten Frist herausstellt, dass das nicht geschieht bzw. geschehen kann, nimmt der ersterfassende Bibliotheksverbund selbst eine Aktualisierung oder Löschung dieser URL vor. Die Titelaufnahmen inkl. URN oder aktualisierter URL werden anschließend an alle anderen Bibliotheksverbände geliefert, damit diese das Ergebnis der Arbeit nachnutzen können, so dass Doppelarbeit vermieden wird. Es ist dann den empfangenden Bibliotheksverbänden überlassen, nach welchem Modus sie die Daten in ihren Katalog einladen: Nur Kategorie für die ID des ersterfassenden Verbun-

des bzw. der DNB (MAB-Feld 001 wird in anderen Datenbanken zu MAB-Feld 026) oder in Kombination mit der Auflagenbezeichnung, dem Erscheinungsjahr, der ISBN etc.

Wenn ein solches Konzept deutschlandweit greift, kann grenzüberschreitend nach Erweiterung Ausschau gehalten werden.