b.i.t.online

BAND 50 INNOVATIV

maLIS

Praxisprojekte 2014

Projektberichte aus dem berufsbegleitenden Masterstudiengang *Bibliotheks- und Informationswissenschaft* der Fachhochschule Köln

Fachhochschule Köln
Cologne University of Applied Sciences

Institut für Informationswissenschaft
Institute of Information Science

2014

b.i.t.online
INNOVATIV

DINGES & FRICK
| Offsetdruck | Digitaldruck | Verlag |

Band 50

b.i.t.online – Innovativ

Band 50

MALIS-Praxisprojekte 2014

Projektberichte aus dem berufsbegleitenden Masterstudiengang
Bibliotheks- und Informationswissenschaft
der Fachhochschule Köln

2014

Verlag: Dinges & Frick GmbH, Wiesbaden

MALIS-Praxisprojekte 2014

Projektberichte aus dem berufsbegleitenden Masterstudiengang
Bibliotheks- und Informationswissenschaft
der Fachhochschule Köln

Herausgegeben
von

Achim Oßwald
Inka Tappenbeck
Haike Meinhardt
Hermann Rösch

2014

Verlag: Dinges & Frick GmbH, Wiesbaden

b.i.t.online – Innovativ

Bibliografische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.
ISBN 978-3-934997-63-9

ISBN 978-3-934997-63-9
ISSN 1615-1577

Redaktion: Dorothee Hofferberth und Susanne Röltgen
Satz: Dorothee Hofferberth
Titelfoto: © iStockfoto.com

Satz und Druck: Dinges & Frick GmbH, Wiesbaden
Printed in Germany

MALIS-Praxisprojekte 2014

Projektberichte aus dem berufsbegleitenden Masterstudiengang Bibliotheks- und Informationswissenschaft der Fachhochschule Köln

Herausgegeben
von

Achim Oßwald
Inka Tappenbeck
Haike Meinhardt
Hermann Rösch

Fachhochschule Köln
Fakultät für Informations- und Kommunikationswissenschaften
Institut für Informationswissenschaft

Interne Kommunikation

Qualifizierung

Ein Konzept für die digitale Langzeitarchivierung des „BIX 2004-2011“

Martin Iordanidis

Abstract

Die Website der „BIX 2004-2011“ ist im Sinne der digitalen Langzeitarchivierung ein komplexes digitales Objekt mit verschiedenen Dateiformaten und Dokumententypen. Für die dauerhafte Verfügbarkeit der Inhalte werden Konzepte der digitalen Langzeitarchivierung vorgestellt und unter Wahrung des früheren Nutzungskontextes technisch angewandt. Abschließend werden basierend auf den gewonnenen Erkenntnissen praktische Empfehlungen für die langfristige Erhaltung der alten BIX-Website formuliert.

In terms of long-term preservation, any website can be considered as a complex digital object with different file formats and document types. This also applies to the now defunct website of the “BIX 2004-2011“, a statistical tool monitoring the German library landscape. The project introduces both retention concepts as well as practical approaches to the preservation of legacy data contained in the discontinued website. The work concludes with recommendations regarding data preparation for the current BIX website.

1. Zielsetzung

Ziel des Projekts war die Erstellung eines Konzepts zur digitalen Langzeitarchivierung der Webpräsenz des Deutschen Bibliotheksindex (BIX)[1] für die Jahre 2004 bis 2011. Zu dem Konzept zählen Überlegungen zur Ermittlung erhaltungswürdiger Eigenschaften des „BIX 2004-2011“, Methoden zur langfristigen Sicherung seiner Inhalte sowie die ersten praktischen Schritte zu deren technischer Umsetzung.[2]

---

1 Vgl. BIX - Der Bibliotheksindex.

2 Das Projekt erfolgte im Rahmen des berufsbegleitenden Weiterbildungsstudiengangs Bibliotheks- und Informationswissenschaft (MALIS) an der Fachhochschule Köln. Betreuer war Dr. Peter Kostädt.

Dem Auftraggeber des Projektes – das hbz[3] als technischer Betreiber des Bibliotheksindex – wurden als Projektergebnis praktische Empfehlungen zur Langzeitarchivierung der Website zur Verfügung gestellt. Das Projekt endete mit einer zusammenfassenden Präsentation der Ergebnisse am 30. September 2013 vor der Berliner Geschäftsstelle des Deutschen Bibliotheksverbandes e. V. (dbv).[4]

2. Ausgangslage

Der Bibliotheksindex (BIX) existiert seit 1999 als Kooperationsprojekt des Deutschen Bibliotheksverbandes (dbv) und der Bertelsmann Stiftung. Seit 2005 wird der BIX vom Deutschen Bibliotheksverband zusammen mit dem Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (hbz) im Rahmen des Kompetenznetzwerks für Bibliotheken (knb)[5] betrieben. Der Bibliotheksindex bietet öffentlichen und wissenschaftlichen Bibliotheken die Möglichkeit eines jährlichen Leistungsvergleiches auf nationaler Ebene. Den Anlass zur Archivierung des alten BIX gab der Umstieg auf eine neue Webpräsenz am 12. Juli 2012. Mit dem Relaunch ist der öffentliche Zugang zur alten BIX-Website (siehe Abb. 1) nicht mehr möglich.

Abb. 1: Startseite der BIX-Website bis Juli 2012

3 Vgl. hbz. BIX - Der Bibliotheksindex.

4 Vgl. Deutscher Bibliotheksverband.

5 Vgl. Bibliotheksportal. knb - Kompetenznetzwerk für Bibliotheken.

Mit dem Redesign der BIX-Website im Sommer 2012 geht auch ein neuer konzeptueller Ansatz einher, der das bis dato erstellte „Ranking“ durch ein „Rating“ ersetzt. Diese Neuerung kann zum einen als eine Loslösung vom Wettbewerbsdenken zwischen den teilnehmenden Bibliotheken interpretiert werden. Zum anderen bringt diese Änderung individuelle Datenprofile mit sich, mit denen sich im Vergleich zum alten BIX repräsentativere Vergleiche ziehen lassen. Der Deutsche Bibliotheksverband schreibt dazu in einem Informationsflyer:

> Statt der bisherigen Gesamtplatzierung im Ranking erhält jede Bibliothek künftig ein individuelles Datenprofil. So wird es möglich, die Positionierung jedes einzelnen Indikators im Verhältnis zu den anderen Bibliotheken der Vergleichsgruppe zu erkennen. Als differenziertes Ergebnis entsteht ein Stärken- / Schwächen-Profil, das auch kleinere Veränderungen zum Vorjahr abbildet.[6]

Die Ranglisten der früheren BIX-Ergebnisse sind mit den Ergebnissen des BIX ab 2012 nicht unmittelbar vergleichbar. Die BIX-Dimensionen, Indikatoren und Vergleichsgruppen sind jedoch weitgehend gleich geblieben. Etwa zeitgleich mit dem Umstieg auf den neuen BIX entand seitens der Betreiber der Wunsch, die Webpräsenz des alten BIX in ihrer bestehenden Form langfristig verfügbar zu halten. Eine Veränderung des alten Datenbestandes war dabei nicht vorgesehen.

Pioniere der Datenmodellierung wie Tsichritzis und Lochovsky sehen in der Beschreibung von Phänomenen der äußeren Welt die Grundlage[7] aller Datenmodelle sowie Daten als deren atomare Einheiten. Sie interpretieren Wahrnehmung als eine Serie von Beschreibungen verschiedenartiger, ggf. verbundener Phänomene*:* “A perception of the world can be regarded as a series of distinct although sometimes related phenomena.“[8] Beschreibungen von Phänomenen, so unvollständig oder unverstanden sie auch sein mögen, werden in diesem Zusammenhang von den Autoren als *Daten* bezeichnet.[9]

Die alte BIX-Website enthält neben den Daten selbst auch Hinweise auf die Interpretation der erhobenen Daten. Neben dem unmittelbaren Zugriff auf die Datenbasis des alten BIX ist ein zweites Argument für die Langzeitarchivierung daher die Erhaltung des hier dokumentierten Kontextes. Er repräsentiert – unter anderem – eine Sicht auf den Datenkorpus, die sich an vergangenen strategischen Zielen des BIX orientiert.

6 Vgl. Neustart für den BIX, S. 2.

7 Tsichritzis; Lochovsky, S. 3.

8 Ebd.

9 Vgl. Iordanidis 2008, S. 29.

3. Problemskizze

In Vorgesprächen mit dem Projektpartner hbz hat sich gezeigt, dass die Portierung des alten BIX auf eine neue technische Plattform sowie die nachhaltige Verfügbarkeit der alten BIX-Webpräsenz zwei verschiedene Ziele sind, zwischen denen Querbezüge existieren. Die Portierung des alten BIX zur weiteren Nutzung im WWW war zum Zeitpunkt der Projektdurchführung technisch noch unabhängig von Maßnahmen der digitalen Langzeitarchivierung. Die dort verwendeten Webtechnologien und Datenbanken sind – noch – verbreitet und erlauben die Portierung auf einen neuen Server. Im Mittelpunkt des Praxisprojektes steht daher ein Konzept zur Langzeitarchivierung des alten BIX, das die *erneute* Portierung des Systems in zeitgemäße Technologien potenziell *jederzeit* wieder ermöglicht.

4. Methodik

4.1 Bestandsaufnahme: die alte BIX-Website als komplexes digitales Objekt

Die Archivierung von Webseiten stellt aus verschiedenen Gründen eine größere Herausforderung dar als die nachhaltige Sicherung von diskreten digitalen Objekten.[10] In der formalen Betrachtung stellt auch die BIX-Website ein solches komplexes digitales Objekt dar.[11] Diese Komplexität ergibt sich v. a. aus

heterogenen Dateiformaten: der alte BIX besteht aus verschiedenen Dateiformaten, darunter HTML-Dateien, Cascading Stylesheets, Javascript-Dateien, PDF-Dokumente, Bilddateien in den Formaten .jpg und .gif;

heterogenen Dokumententypen: in ihrer Rolle als Dokumententypen müssen die technischen Dateiformate aus einer anderen, eher editorischen Perspektive betrachtet werden. Unter den vorgefundenen PDF-Dokumenten und HTML-Seiten befinden sich

- Portraits von teilnehmenden BIX-Bibliotheken
- Vortragsskripte
- statistische Ergebnisse aus einer definierten Anwendersicht
- kritische Auseinandersetzungen mit dem Bibliotheksindex
- Interviews mit Anwendern

und damit wichtige kontextuelle Informationen rund um den BIX.

10 Vgl. Hennies 2010, S. 70 ff.

11 Die Literatur zur *significant properties* bezieht sich i. d. R. auf die Eigenschaften eines diskreten digitalen Objekts. Im Kontext dieser Arbeit wird diese Betrachtung jedoch auf die gesamte BIX-Website im Sinne eines *komplexen digitalen Objektes* angewandt.

„Deep Web“-Inhalten: Die Datenbasis des alten BIX ist in einer MySQL-Datenbank hinterlegt, welche über PHP-Skripte angesprochen wird. Weiterhin werden die meisten redaktionellen Inhalte der Website mit dem Enterprise-Content-Management-System TYPO3[12] verwaltet. Aus diesen Voraussetzungen ergibt sich, dass verschiedene „Pre-Ingest“-Verfahren angewendet werden müssen, um die erhaltungswürdigen Eigenschaften des alten BIX zu sichern. Erleichternd wirkt sich dabei aus, dass durch den Projektpartner hbz ein direkter Zugriff auf die technische Plattform möglich ist. Die Langzeitarchivierung von Datenbankinhalten wäre ansonsten nicht zu bewerkstelligen.

Aus dieser ersten Analyse ergeben sich zwei verschieden Sammlungsverfahren für die Sicherung der Daten: zum einen das serverseitig gesteuerte *Webharvesting*[13] für Inhalte, die unmittelbar durch einen Browser interpretierbar sind. Zum anderen ein Export der BIX-Daten zusammen mit der Datenbankstruktur in ein archivtaugliches Format.

4.2 Vorüberlegungen zu signifikanten Eigenschaften des alten BIX

Leitfragen bei der Ermittlung von *significant properties* im Kontext der Langzeitarchivierung lauten: Was interessiert mutmaßlich, jetzt und zukünftig, die Fachcommunity? Wer ist überhaupt die Fachcommunity und welche könnte zukünftig dazuzählen? Das britische Kompetenzzentrum Jisc[14] definiert signifikante Eigenschaften als die Charakteristika eines digitalen Objektes, die sich auf

- das *Erscheinungsbild*
- das *Verhalten*
- die *Qualität*
- die *Nutzbarkeit*

eines des digitalen Objektes auswirken. Sie können gruppiert werden nach ihrem *Inhalt*, einem mittels Metadaten beschriebenen *Kontext*, nach ihrem visuellen *Erscheinungsbild* („Layout“) und ihrer *Funktionalität*. Signifikante Eigenschaften sollten erhalten werden, damit die tatsächlich relevanten Aspekte digitaler Daten langfristig zugänglich und nutzbar bleiben. Die Ermittlung der Signifikanz einer Eigenschaft ist in den meisten Fällen subjektiv und damit direkt an das Nutzerverhalten der Zielgruppe gekoppelt. Da im Kontext dieser Arbeit keine Vorabanalyse mit den Nutzergruppen durchgeführt werden konnte, wird im Folgenden eine Liste mutmaßlich signifikanter Eigenschaften in der Reihenfolge ihrer Wichtigkeit vorgeschlagen:

1. Erhaltung der Datenbasis „BIX 2004-2011“
2. Erhaltung der Datenbankstruktur

12 Vgl. TYPO3.

13 Liegmann 2006, S. 42 ff.

14 Vgl. Jisc 2008.

3. Erhaltung des Kontextes; hier: textuelle Inhalte und Bilder
4. Erhaltung der logischen Struktur, hier: der Website
5. Erhaltung des Layouts

Im folgenden Kapitel werden Risiken und Lösungsstrategien der digitalen Langzeitarchivierung vorgestellt und auf die Problemskizze angewandt.

4.3 Risikoebenen und Lösungsstrategien der Langzeitarchivierung

Praktisch wird unter dem Begriff „Langzeit" derzeit ein technisch überschaubarer Zeitraum von fünf bis zehn Jahren verstanden. Dennoch ist das prinzipielle Ziel der digitalen Langzeitarchivierung die *dauerhafte* Nutzbarkeit von digitalen Ressourcen, auch wenn im Fall des alten BIX derzeit keine akute Gefährdung der digitalen Inhalte vorliegt. In den folgenden Abschnitten werden die Risikoebenen im Umgang mit digitalen Ressourcen sowie mögliche Lösungsstrategien der digitalen Langzeitarchivierung überblicksartig erläutert. Dies dient als theoretische Fundierung der Entscheidungen, die für das Langzeitarchivierungskonzept „BIX 2004-2011" zu treffen sind. Hierbei werden nur die Risikoebenen und Lösungsstrategien herausgegriffen, die auch auf den hier behandelten Praxisfall anwendbar sind.

Abb. 2: Schematische Darstellung von Risiken der digitalen Datenhaltung[15]

15 Vgl. Iordanidis 2013, S. 2 ff.

4.4 Risikoebene Bitstream-Verluste

Die grundlegendste technische Ebene aller Erhaltungsstrategien besteht in der physischen Erhaltung des Datenstroms (engl. Bitstream). Ein Bitstream ist eine Sequenz von Bits von unbestimmter Länge in zeitlicher Abfolge, die in weitere logische Abfolgen untergliedert ist. Die elementarste Einheit eines Bitstreams sind Binärziffern, die in Computern und auf Speichermedien als „0“ und „1“ codiert sind. Eine ausführlichere Beschreibung von Bitstreams findet sich bei Rothenberg.[16] Auf der Bitstream-Ebene wirken mehrere Risiken ein. Zum einen können während des Transfers Datenpakete unvollständig übertragen werden oder andere Übermittlungsfehler auftreten. Zum anderen können Bitstreams jederzeit und ohne erkennbare Außeneinwirkung zerfallen („bit rot“) und damit die Lesbarkeit der enthaltenen Information einschränken oder unmöglich machen:

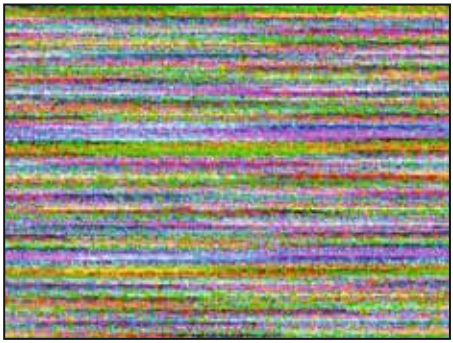

Abb. 3 Schaden durch Bitstream-Verluste[17]

Abb. 4a/b: Informationsverlust durch ein fehlerhaftes Byte von 360.000 Byte[18]

16 Vgl. Rothenberg 2010, S. 5 ff.

17 Vgl. http://commons.wikimedia.org/wiki/File:JPEG_Corruption.jpg

18 Vgl. Thaller 2009, S. 21 f.

4.5 Lösungsstrategien Bitstream Preservation

Die Firma Portico zählt das Einstellen eines Dienstes zu den so genannten *trigger events*,[19] die Maßnahmen der digitalen Langzeitarchivierung auslösen können bzw. sollten. Ein solcher *trigger event* ist mit dem Abschalten des alten BIX im Juli 2012 eingetreten. Während die Webarchivierung üblicherweise vor dem Eintreten dieses Ereignisses geschehen sollte, ist im Falle des alten BIX ein kontrolliertes Einsammeln der Inhalte auch nach der Abschaltung möglich gewesen.

Webharvesting mit HTTrack

In Form eines unspezifischen Verfahrens ist zunächst die gesamte Domain geharvestet worden. Dies geschah durch den Einsatz des Webcrawlers HTTrack,[20] der im hbz u. a. als Service an die Repositorysoftware DigiTool angebunden ist.

Der Harvestingvorgang beschränkte sich auf die unmittelbar sichtbaren Inhalte der BIX-Website und erzeugte ein 142,4 Megabyte großes Webarchiv. Die Sichtung der im Browser aufrufbaren Dateien ergab, dass die lokale Archivkopie vollständig und alle nicht aus der BIX-Datenbank erzeugten Inhalte über die Navigation erreichbar waren.

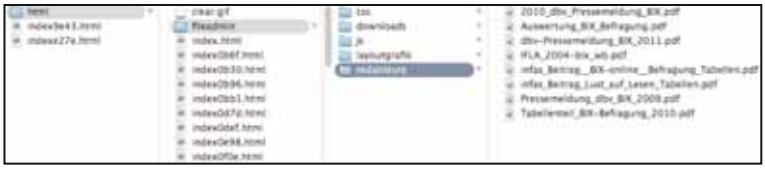

Abb. 5: Systeminterne Identifier für HTML-Dokumente

Jedoch ergab sich aus der Verwendung des Content-Management-Systems TYPO3, dass die Unterseiten systeminterne Identifier besitzen statt „sprechende Dateinamen". Weiterhin bildet die Verzeichnisstruktur der Unterseiten nicht die Binnenstruktur der Website ab. Dies ist einer der Gründe dafür, warum das Layout der alten BIX-Website als signifikant bewertet werden könnte: die Struktur der Website wäre über das Layout bzw. die Hauptnavigation am leichtesten rekonstruierbar.

Zur Erhaltung der Websitestruktur bieten sich zwei denkbare Optionen. Zum einen wäre die Abbildung der Struktur in einem METS-Dokument[21] möglich. METS erlaubt als so genanntes *Containerformat* die Abbildung der Datenstruktur in einer *struct map* und ermöglicht darüber hinaus die Integration technischer Metadaten für jedes einzelne digitale Objekt. Dieser Weg wäre jedoch unverhältnismäßig aufwändig, zumal sich in

19 Portico 2007.

20 HTTrack 2013.

21 Vgl. Wikipedia: METS - Metadata Encoding & Transmission Standard.

dem Webarchiv keine seltenen bzw. undokumentierten Dateiformate befinden. Daher wurde entschieden, das gesamte Webarchiv mit seiner bestehenden Verzeichnisstruktur in einer .zip-Datei zu komprimieren. Das .zip-Format wird von der Library of Congress als weitgehend geeignet für die digitale Langzeitarchivierung eingeschätzt.[22] Zwar hat es einen proprietären Ursprung, ist aber offen dokumentiert und sehr weit verbreitet.

Aufgrund des sehr hohen Textanteils ergab sich nach der Kompression des BIX-Webarchivs eine im Vergleich geringe Dateigröße von 28,8 Megabyte.

Erstellung von Prüfsummen und mehrfach redundante Speicherung

Zu den gängigen Strategien der Bitstream-Erhaltung zählt die Erstellung von Prüfsummen beim Transfer von Daten sowie die verteilte redundante Datenspeicherung – d. h. eine mehrfache Speicherung identischer Daten an räumlich getrennten Orten. Während Prüfsummenverfahren eine relativ einfache Maßnahme zur Gewährleistung der Datenintegrität darstellen, ist mit der redundanten Datenspeicherung ein höherer infrastruktureller Aufwand verbunden. Die von Rechenzentren als Standarddienstleistung gängige Erstellung von Backups kann dabei technisch nicht mit digitaler Langzeitarchivierung gleichgesetzt werden. Zwar greifen Backups als kurz- bzw. mittelfristige Sicherungsmaßnahme mit der digitalen Langzeitarchivierung auf Bitstream-Ebene ineinander; die über lange Zeiträume zu beobachtenden Technologiesprünge von Dateiformaten, Software und Betriebssystemen werden allein durch Backups jedoch nicht berücksichtigt. Technologiesprünge stellen somit ein breites Problemfeld in der digitalen Langzeitarchivierung dar. Ein vor allem in der Bibliothekswelt verbreiteter Ansatz zur Bitstream Preservation ist die von der Standford University entwickelte Open-Source-Software LOCKSS (Lots Of Copies Keep Stuff Safe).[23] Mit niedrigen technischen Hürden und vergleichsweise preiswerter Hardware kann mit LOCKSS ein regional, national oder global verteiltes Speichernetzwerk aufgebaut werden, das auf Veränderungen von Bitstreams dynamisch reagiert. Datenströme werden in Echtzeit auf ihre Integrität geprüft und bei einer festgestellten Datenkorruption durch einen intakten, andernorts gespeicherten Datenstrom ersetzt.

Als Vorbereitung für die redundante Speicherung in einem LOCKSS-Netzwerk wurde mit der Terminalanwendung von Mac OS X folgende MD5-Prüfsumme[24] für das Webarchiv erstellt:

ec67f219777f2b1663dbf16a664989af

22 Vgl: Library of Congress. Digital Preservation: Sustainability of Digital Formats: ZIP File Format.

23 Vgl. LOCKSS.

24 Vgl. Wikipedia: Message-Digest_Algorithm_5.

MD5 gilt unter kryptographischen Gesichtspunkten inzwischen nicht mehr als sichere Verschlüsselungsmethode für den Datentransfer im World Wide Web.[25] In einem kontrollierten und technisch geschlossenen System wie LOCKSS überwiegen jedoch die praktischen Vorteile dieser verbreiteten Technologie. LOCKSS selbst ist in der Lage, neben MD5 auch das verwandte Verfahren SHA-1[26] zu verarbeiten.

Das hbz betreibt zwar eine eigene „LOCKSS-Box" innerhalb eines deutschen LOCKSS-Netzwerkes, die Einspielung des BIX-Webarchivs wurde aber im Zuge des Praxisprojektes nicht vorgenommen. Für diesen Schritt müsste zunächst die explizite Zustimmung der Rechteinhaber eingeholt werden, da auch eine Speicherung zu Archivzwecken als eine Vervielfältigung im Sinne des derzeit geltenden Urheberrechts gilt.[27]

4.6 Risikoebene Gefährdete Formate

Die Informationstechnologie hat in den letzten Jahrzehnten eine immense Vielfalt von *Dateiformaten* und Software hervorgebracht. Schätzungen gehen von aktuell über 550 Dateiformaten aus, denen über 10.000 historisch gewachsene Dateiformate gegenüberstehen.[28] Zahllose Formate sind durch proprietäre Herstellerinteressen geprägt und gelten, unzureichend oder auch gar nicht dokumentiert, als *„black box"*, deren zukünftige Nutzbarkeit in Frage zu stellen ist. Weitere Risiken ergeben sich aus den Abhängigkeiten zwischen Dateiformaten und Software, zwischen Software und Betriebssystemen sowie zwischen Betriebssystemen und Hardware.

Es verwundert vor diesem Hintergrund nicht, dass frühe Bestrebungen hin zur Nachhaltigkeit von digitalen Daten sich auf der Ebene der Dateiformate abgespielt haben. Aus technischer Sicht ist ein Dateiformat eine genau definitere Ordnung von Bitfolgen, die idealerweise im Rahmen einer *Formatspezifikation* dokumentiert ist. Mit Hilfe einer transparenten Formatspezifikation kann die Struktur und Beschaffenheit einer Datei nachvollzogen und von Computern interpretiert werden. Bei proprietären Dateiformaten ist die Spezifikation oft nicht bekannt, während sie bei *offenen Formaten* frei zugänglich ist. Dokumentierte Spezifikationen und *standardisierte Formate* sind für die Langzeitarchivierung digitaler Daten essentiell und stellen das wichtigste Kriterium *archivtauglicher Dateiformate* dar. Intransparente Dateiformate erschweren Technologiewechsel dagegen erheblich. Verschiedene Nutzercommunitys haben für ihre Domäne jeweils archivtaugliche Dateiformate definiert, die abseits ihrer heterogenen Einsatzzwecke und Eigenschaften einige gemeinsame Kriterien aufweisen. So sollten archivtaugliche Formate möglichst verlustfrei sein, technische Metadaten enthalten, transparent dokumentiert

25 Ebd.

26 Vgl. Wikipedia: Secure Hash Algorithm.

27 Sietmann 2011.

28 Vgl. Fileformat.info.

und nicht (oder nicht mehr) proprietär sein.[29] Eine Auswahl archivtauglicher Formate findet sich u. a. beim Schweizerischen Bundesarchiv.[30]

4.7 Lösungsstrategie Formatmigration

4.7.1 Umgang mit HTML-Seiten

Die kontextuellen Informationen des BIX liegen hauptsächlich in den Formaten .html und PDF vor. Verfolgt man einmal die Entstehungsgeschichte von HTML zurück in ihre Anfangstage, werden Parallelen zu den heutigen Bestrebungen der digitalen Langzeitarchivierung sichtbar. Versteht man HTML als eine Anwendung der Standard Generalized Markup Language (SGML) für das World Wide Web,[31] dann finden sich in auch in HTML originäre Prinzipien der Langzeitarchivierung wieder. Dazu zählt die Auszeichnung von (hier: textuellen) Daten mit Metadaten, die neben ihrer Funktion als Formatierungsanweisung für den Browser auch semantische Informationen zu den Textfragmenten enthalten. Beispiele dafür sind die Relevanz (Formatierung als Überschrift), die Herkunft (Formatierung als Zitat) oder die Relation zu anderen Texten (Formatierung als Hyperlink).

Bei der Analyse des Webarchivs fiel weiterhin auf, dass im Unterbereich „Ergebnisse“ Daten aus der BIX-Datenbank innerhalb der HTML-Seiten mitarchiviert wurden. Die ist begründet durch „hart-codierte“ Datenbankanfragen an die BIX-Datenbank, die mitsamt Ergebnissen im Zuge des Harvestings eingesammelt wurden.

```
(...)
<TR bgColor=#ffffff><TD class=C_H2 colSpan=2 vAlign=top><BR>
<Ahref="index9c91.html?nID=21&amp;year=2011&amp;lnr=1463&amp;num=2">
Ausgangsdaten </A></TD><td align= " right"  valign= "bottom">
<A href="index9c91.html?nID=21&amp;year=2011&amp;lnr=1463&amp;num=2">
(...)
href="index0cbe.html?nID=21&amp;year=2011&amp;lnr=1463&amp;num=3">
Ergebnisse nach Einwohnergr&ouml;&szlig;enklassen </A></TD>
<td align="right" valign="bottom">
<A href="index0cbe.html?nID=21&amp;year=2011&amp;lnr=1463&amp;num=3">
```

29 Vgl. nestor Handbuch 2010.

30 Standards für die Archivierung digitaler Unterlagen 2007, S. 5 f.

31 Vgl. Iordanidis 2006, S. 24 ff.

Die HTML-Seiten wurden in ihrem Format belassen, da sie ausgiebig dokumentiert, mittelfristig von Browsern interpretierbar und langfristig als Klartextdateien mit einfachen Editoren lesbar sind.

PDF/A-Migration

Für die insgesamt 39 PDF-Dokumente des Webarchivs wurde eine Formatmigration in das archivtaugliche Format PDF/A-1b vorgenommen. Mit diesem PDF/A-Level wird bei einer gelungenen Formatmigration die visuelle Reproduzierbarkeit garantiert, indem alle für die Reproduktion nötigen Schriften in das PDF/A-Dokument eingebettet werden. Für die Migration wurde das Tool PDF/A-Pilot der Firma callas in der Version 4.3.186 eingesetzt, welches beim hbz als Migrationsservice genutzt wird. Zusammen mit den Formatmigrationen wurden Konvertierungsreports angefertigt, in denen Korrekturen bzw. Probleme bei der Konvertierung dokumentiert sind:

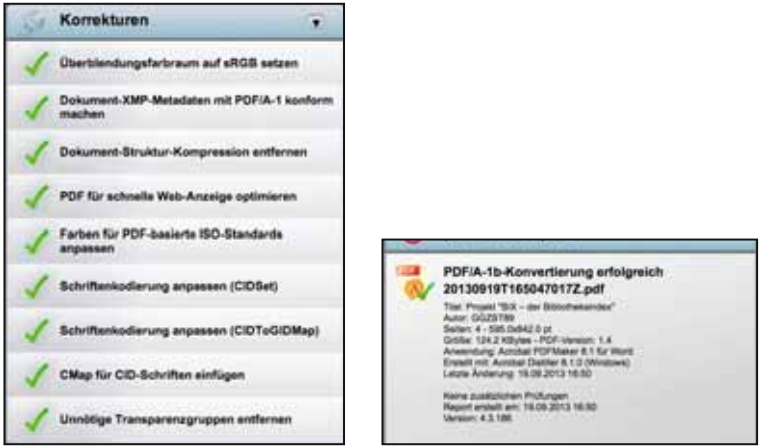

Abb. 6 a/b: Grafische Migrationsreports des callas-PDF/A-Pilots

Von den 39 PDF-Dokumenten konnten auf diese Weise 36 erfolgreich nach PDF/A migriert werden, was einer Quote von ca. 92 % entspricht. Perspektivisch sind auch Konvertierungsreports im XML-Format erstellbar, die für den Einsatz in Langzeitarchivierungssystemen nützlich sein können. Da in dem Praxisprojekt das intellektuelle Verständnis der Datenbasis PDF im Vordergrund stand, wurden nur die visuellen Konvertierungsreports zusammen mit den PDF/A-Dokumenten erstellt.

Datenbankmigration

Die statistischen Daten des alten BIX sind in einer MySQL-Datenbank gespeichert. MySQL gilt als die populärste Open-Source-Software für Datenbankanwendungen und ist sehr gut dokumentiert. Diese beiden Faktoren sind prinzipiell zuträglich für die Langzeitarchivierung. Dennoch bestehen Abhängigkeiten zwischen dem Datenkorpus und der verwendeten Software, in weiterer Instanz auch zwischen dem verwendeten Betriebssystem und letztlich auch zur Hardware. Hinsichtlich der Archivtauglichkeit sind verschiedene Formate für Datenbankbestände denkbar. Das Schweizerischen Bundesarchiv empfiehlt das Format SIARD (Software Independent Archiving of Relational Databases),[32] welches mit Hilfe eines Software-Toolkits auch aus MySQL-Daten ein Archivpaket erstellen kann. Ein Blick in die Formatspezifikation von SIARD offenbart, dass SIARD auf den ISO-Normen Unicode und XML basiert.

Datenbankexporte in das XML-Format haben neben der Verwendung eines verbreiteten W3C-Standards noch mindestens zwei weitere entscheidende Vorteile: zum einen können Datenstrukturen aus relationalen Datenbanken in XML leicht abgebildet werden. Zum anderen ist XML weitgehend unabhängig von interpretierender Software und notfalls auch als Klartext lesbar. Der große Nachteil von XML-Datenbanken ist die schlechte Performanz bei Datenabfragen. Da im Falle des BIX kein regelmäßiger Datenbankzugriff, sondern ein zuverlässiges Archivformat gefragt ist, wurde für das Praxisprojekt dieser niedrigschwellige Weg gewählt.

Der XML-Export der BIX-Datenbank ergab eine Datei von ca. 292 Megabyte Größe. Komprimiert als .zip-Datei reduzierte sich die Dateigröße bedingt durch die große Redundanz in Markup-Sprachen auf 8,8 Megabyte. Diese geringe Größe macht den gesamten Datenbestand ausgesprochen transferfähig und damit auch für den Einsatz in redundanten Speichersysteme wie LOCKSS denkbar. Als Nächstes wurde auch für die BIX-Datenbank eine MD5-Prüfsumme erstellt und zusammen mit der gezippten XML-Datei lokal im hbz abgespeichert.

4.8 Lösungsstrategie Rechtssicherheit

Die nächsten logisch erscheinenden Schritte zur Langzeitarchivierung des alten BIX sind die Sicherung in einem redundanten Speichernetzwerk sowie eine aktive Förderung der Nachnutzung der BIX-Daten. Brewster Kahle, der Gründer des Internet Archive, fasst die Nutzung von Daten mit dem Satz “Access drives preservation[33] prägnant zusammen. Diesem zentralen Überlebensfaktor digitaler Daten stehen jedoch auf der Seite der Gesetzgebung ein veraltetes Urheberrecht sowie unklare Rechtslage auf Seiten der Urheber entgegen. Betreiber könnten durch eine offene Lizenz wie z. B. CC0 (Crea-

32 Datenbankarchivierung: SIARD Suite 2012.

33 Cieplak-Mayr von Baldegg 2013.

tive Commons 0)[34] Rechtssicherheit in Bezug auf die Nutzung ihrer Daten schaffen. Weiterhin würde eine Öffnung der Daten für Linked-Open-Data-Anwendungen[35] das Nachnutzungspotenzial erheblich erhöhen. Der „Berliner Appell zum Erhalt des digitalen Kulturerbes" adressiert diese Problematik in folgendem Punkt:

> (...) 4. Recht
>
> Der derzeitige Rechtsrahmen behindert vielfach die digitale Langzeitarchivierung. Es müssen eindeutige und verlässliche rechtliche Rahmenbedingungen für die digitale Langzeitarchivierung in all ihren Aspekten geschaffen werden.[36]

Die Öffnung von Datenbeständen im Kontext von Linked Open Data kann erheblichen Mehrwert für verschiedene Zielgruppen schaffen und nützt letztlich auch den Urhebern der Daten. Ein denkbares Szenario wäre die Integration der BIX-Daten in andere statistische Werkzeuge wie den im September 2013 in Spiegel Online veröffentlichte „Wahlaltlas":

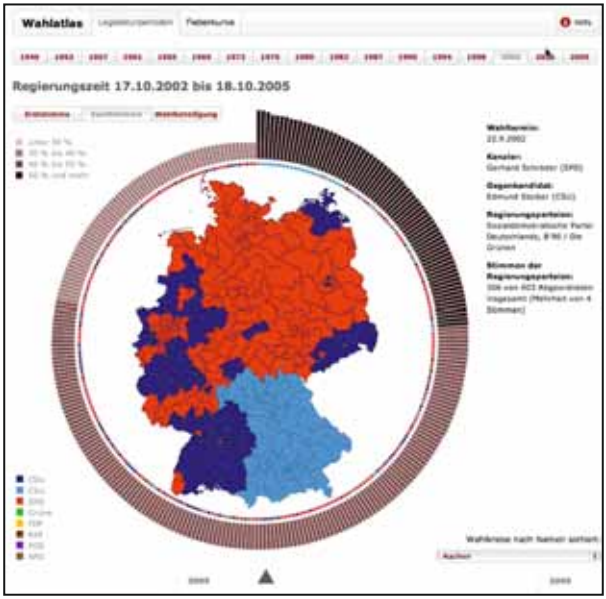

Abb. 7: Screenshot des „Wahlatlas" auf Spiegel Online im September 2013

Auf diese Weise ließen sich beispielsweise die Leistungsprofile von Bibliotheken zu den Legislaturperioden verschiedener Bundesregierungen in Beziehung setzen.

34 Vgl. Wikipedia: Creative Commons.

35 Vgl. Linked Open Data 2012.

36 Vgl. Berliner Appell zum Erhalt des digitalen Kulturerbes 2013.

5. Zusammenfassende Empfehlungen

Abschließend lassen sich basierend auf den Erkenntnissen des Praxisprojektes folgende Empfehlungen für die praktische Langzeitarchivierung des „BIX 2004-2011“ formulieren:

- Entscheiden: Welche Eigenschaften des BIX sind erhaltungswürdig, welche eher nicht?
- Einsammeln der Inhalte mittels Webharvesting / Sicherung vom Webserver
- Archivtaugliche Formate: Export der BIX-Datenbank, z. B. in XML
- Archivtaugliche Formate: Bestimmung des geeigneten PDF/A-Conformance-Levels und Konvertierung in PDF/A-(x)
- Bitstream Preservation: Speicherung der BIX-Datenbank und ggf. Prüfsumme in einem redundanten Speichernetzwerk (LOCKSS).
- Beibehaltung der Originaldatenbank und -PDFs
- Rechtssicherheit schaffen, z. B. durch offene Lizenz
- Aktive / passive Förderung der freien Nachnutzung

Nur in wenigen Fällen erlaubt eine direkte Zugriffsmöglichkeit auf Datenbasen die unmittelbare Migration von Datenbankinhalten so wie im vorliegenden Projekt. Für die Betreiber von datenbankgestützten Websites lassen sich jedoch trotzdem einige Fingerzeige ableiten, um Inhalte langfristig einfacher archivierbar zu machen. Dabei sei besonders auf den Einsatz nicht-proprietärer Objektformate sowie die Dokumentation der Datenbankstruktur verwiesen. Übertragen auf die gängigen Herausforderungen des Websiteharvestings ist eine Reflektion über sämtliche zu erhaltenden Eigenschaften einer Website sowie deren kontextuelle Ebene unabdingbar. Sie grenzt für Websitebetreiber und Spezialisten der Langzeitarchivierung Handlungsfelder ab und bildet die Grundlage weiterer technischer Schritte – bereits vor und während des Hostings.

Die Dokumentation inhaltlicher und kontextueller Informationen geht dabei über einen reinen Selbstzweck hinaus. Als semantische Brücke zu Nachweisinstrumenten leisten sie der weiteren Benutzung der Inhalte Vorschub, womit sich wiederum finanzielle und andere infrastrukturelle Aufwände der Langzeitarchivierung rechtfertigen lassen.

Im Nachgang des Projekts wird auch die Langzeitarchivierung des aktuellen BIX beim Träger der Deutschen Bibliotheksstatistik, dem Deutschen Bibliotheksverband e. V., auf ihre Machbarkeit hin untersucht.

Martin Iordanidis

studierte Musikwissenschaft und Historisch-Kulturwissenschaftliche Informationsverarbeitung an der Universität zu Köln und arbeitet seit 2008 im Bereich Digitaler Langzeitarchivierung. Er ist seit 2004 im Hochschulbibliothekszentrum des Landes NRW in der Gruppe Publikationssysteme beschäftigt und absolviert seit 2012 den berufsbegleitenden Masterstudiengang Bibliotheks- und Informationswissenschaft an der Fachhochschule Köln.

Kontakt: martin.iordanidis@googlemail.com

Abbildungsverzeichnis

Literatur- und Quellenverzeichnis

Abrufdatum der Internetdokumente ist der 01.09.2013. Im Einzelfall abweichende Abrufdaten sind angegeben.

Archivtaugliche Dateiformate. Standards für die Archivierung Digitaler Unterlagen. 2007. In: Schweizerisches Bundesarchiv Ressort Innovation und Erhaltung. http://www.bar.admin.ch/dienstleistungen/00895/00897/index

Berliner Appell zum Erhalt des digitalen Kulturerbes. 2013. http://www.berliner-appell.org

Bertelsmann Stiftung. BIX - Der Bibliotheksindex. http://www.bertelsmann-stiftung.de/cps/rde/xchg/SID-8776D537-23A84135/bst/hs.xsl/5484.htm

Bibliotheksportal. knb - Kompetenznetzwerk für Bibliotheken. http://www.bibliotheksportal.de/wir-ueber-uns/kompetenznetzwerk.html

BIX - Der Bibliotheksindex. http://www.bix-bibliotheksindex.de. (URL am 01.05.2013, inzwischen nicht mehr öffentlich aufrufbar).

Cieplak-Mayr von Baldegg, Kasia 2013: Inside the Internet Archive. In: The Atlantic. http://www.theatlantic.com/technology/archive/2013/05/inside-the-internet-archive/275610

Datenbankarchivierung: SIARD Suite. 2012. In: Schweizerisches Bundesarchiv Ressort Innovation und Erhaltung. http://www.bar.admin.ch/dienstleistungen/00823/00825/

Deutscher Bibliotheksverband. http://www.bibliotheksverband.de

Fileformat.info. The Digital Rosetta Stone. http://www.fileformat.info

hbz. BIX - Der Bibliotheksindex. http://www.hbz-nrw.de/angebote/dbs/bix

Hennies, Markus 2010: Webarchivierung in der Praxis. In: Langzeit-archivierung komplexer Objekte. Vortragsskript der nestor summer school 2010 (unveröffentlicht).

Iordanidis, Martin 2008: XML Schema für Daten und Metadaten im Bereich DigitalerBibliotheken. Saarbrücken: VDM.

Jisc. 2008.The significant properties of digital objects. http://www.jisc.ac.uk/whatwedo/programmes/preservation/2008sigprops

Library of Congress. Digital Preservation: Sustainability of Digital Formats: ZIP File Format. http://www.digitalpreservation.gov/formats/fdd/fdd000354.shtml

Liegmann, Hans 2006: Web-Archivierung zur Langzeiterhaltung von Internet-Dokumenten. In: nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Neuroth, Heike [u.a.] (Hrsg.). Boizenburg: Hülsbusch, Kap. 17:88-17:103. http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_146.pdf

Linked Open Data. 2012. In: Hochschulbibliothekszentrum des Landes new (hbz). http://opendata.hbz-nrw.de sowie http://www.hbz-nrw.de/dokumentencenter/produkte/lod/aktuell

LOCKSS. Lots Of Copies Keep Stuff Safe: What Is LOCKSS? http://www.lockss.org/about/what-is-locks

nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarichivierung. 2010. Version 2.3. Neuroth, Heike [u.a.] (Hrsg.). Boizenburg: Hülsbusch. http://nestor.sub.uni-goettingen.de/handbuch/nestor-handbuch_23.pdf Göttingen

Neustart für den BIX. 2012 erhält der BIX ein neues Konzept. http://www.oebib.de/fileadmin/redaktion/meldungen/2011_2/10_BIX_Flyer.pdf

Portico. 2007. Trigger Events. http://www.portico.org/digital-preservation/services/reliable-access/trigger-events

Rothenberg, Jeff 1999: Ensuring the Longevity of Digital Information. http://www.clir.org/pubs/archives/ensuring.pdf (URL am 01.09.2013).

Sietmann, Richard. 2011. „Klarheit für die digitale Langzeitarchivierung im Urheberrecht gefordert". In: Heise.de http://www.heise.de/newsticker/meldung/Klarheit-fuer-die-digitale-Langzeitarchivierung-im-Urheberrecht-gefordert-1196752.html

Thaller, Manfred. 2009. The eXtensible Characterisation Languages - XCL. Hamburg: Kovac.

Tsichritzis, Dionysos ; Lochovsky, Frederick 1982.: Data Models. Englewood Cliffs, N.J.: Prentice Hall.

TYPO3. http://typo3.org

Wikimedia Commons: JPEG Corruption.jpg (23.04.2014).

Wikipedia: Creative Commons. http://de.wikipedia.org/wiki/Creative_Commons#CC0

Wikipedia: HTTrack. http://de.wikipedia.org/wiki/HTTrack

Wikipedia: Metadata Encoding & Transmission Standard. http://de.wikipedia.org/wiki/Metadata_Encoding_%26_Transmission_Standard

Wikipedia: Message-Digest_Algorithm_5. http://de.wikipedia.org/wiki/Message-Digest_Algorithm_5

Wikipedia: Prüfsumme. http://de.wikipedia.org/wiki/Pr%C3%BCfsumme

Wikipedia: Secure Hash Algorithm. http://de.wikipedia.org/wiki/Secure_Hash_Algorithm