

Antje Schlieter

Archiving Websites – Archivierungskonzept für das Intranet der Dresdner Bank AG¹

Webseiten und Websites, eine neue Quellgattung, die auch im Zuständigkeitsbereich des Archivs liegt, rückt in den Blick archivischer Überlieferungsbildung.

Der Archivierung von Websites widmete ich meine Diplomarbeit an der Fachhochschule in Potsdam, die ich im Frühjahr diesen Jahres schrieb. Unter dem Titel „Archiving Websites – Archivierungskonzept für das Intranet der Dresdner Bank AG“ stellte ich einen Archivierungsablauf für das Intranet des Unternehmens vor. Die Ergebnisse der Diplomarbeit bilden die Grundlage des Projekts „Archivierung des Intranets der Dresdner Bank“, das seit September im Historischen Archiv umgesetzt wird. In Rahmen der Diplomarbeit konnten nicht alle Fragen und Probleme gelöst werden, die in diesem Zusammenhang aufgetaucht sind. Die Lösungsansätze jetzt bei der Durchführung erarbeitet.²

Der vorliegende Beitrag gibt eine Zusammenfassung der Diplomarbeit. Auf den aktuellen Stand des Projektes im Historischen Archiv der Dresdner Bank wird nicht weiter eingegangen. Die Diplomarbeit umfaßt die Ausgangssituation des Projektes, die Formulierung der Ziele für den Archivierungsprozeß, die Vorgehensweisen, Allgemeines zur Archivierung von Websites, eine Analyse des Mediums, den entwickelten Archivierungsablauf, die Bewertung von Websites und des Intranets, den Spiegelungsprozeß³ sowie Maßnahmen zur Sicherung und Erhaltung des Intranet-Archivs.

Ausgangssituation

Das Intranet der Dresdner Bank besteht seit 1996 und hat sich seit Inbetriebnahme rasant weiterentwickelt. Es umfaßt nach einer Statistik von Mai 2003 mehr als 800.000 einzelne Webseiten, die überwiegend statisch sind. Das Intranet entspricht

¹ Zusammenfassung der Diplomarbeit zur Erlangung des Titels Dipl.-Archivarin an der Fachschule Potsdam (vom Juli 2003).

² Ungeklärt blieb die Entwicklung und Finanzierung eines Migrationstools, die Benutzungs- und Präsentationsform sowie das Datenhaltungssystem für die hierarchisch strukturierten Daten. Die Anpassung der Indexierungssoftware an das Intranet-Archiv wird im Jahr 2004 realisiert.

³ Auf das Kapitel „Spiegelung“ wird nicht weiter eingegangen. Die Arbeitsweise der Software wird im Archivierungsablauf kurz erklärt.

zur Zeit noch der Charakteristik einer umfassenden Unternehmenspublikation mit archivwürdigen Informationen.⁴

Das Intranet enthält archivwürdige Informationen, die Bestandteile von Geschäftsprozessen im Unternehmen sind und zum Teil nur noch online vorliegen. Dazu zählen u.a. interne Rundschreiben, aktuelle Geschäftsinformationen, Nachrichten, Organigramme oder Protokolle von Arbeitskreissitzungen.

Schließlich muß sich der Archivar auch die Frage stellen, ob nicht historisches Interesse zukünftiger Generationen an der neuen Quellengattung besteht. Diese Frage ist aus meiner Sicht mit Ja zu beantworten.

Bei der Überlieferungsbildung wird der Archivar mit verschiedenen Problemen konfrontiert. Zum einen ist es Flüchtigkeit der Web-Contents und die rasante technische Entwicklung. Zum anderen ist nicht gewährleistet, daß die Informationen nach Ablauf der gesetzlichen Aufbewahrungsfrist in das Archiv vollständig, lesbar und interpretierbar übernommen werden. Ferner existiert das Problem, rückwirkend die Informationen in den ursprünglichen Zusammenhang darzustellen. Technisch ist dies nur mit einem sehr hohen Aufwand verbunden, das noch keine Garantie für eine authentische Überlieferung ist. Deshalb muß das Archiv bereits während der Publikationsphase im Intranet agieren!

Ziele

Die Bemühungen für die dauerhafte Aufbewahrung des Intranets bestanden seitens des Historischen Archivs seit dem Frühjahr/Sommer 2002. Es existierten bisher jedoch keine genauen Vorstellungen, wie dieses Vorhaben realisiert werden sollte. Dazu soll diese Diplomarbeit einen entscheidenden Teil beitragen, um Methoden und Lösungswege aufzuzeichnen.

Ziel des Archivs ist es, daß die Charakteristik des Mediums erhalten bleibt, so daß eine Navigation und Recherche innerhalb des archivierten Teils des Intranets möglich ist. Die archivierten Aufzeichnungen müssen authentisch, zuverlässig, lesbar und interpretierbar sein. Das heißt, es muß gewährleistet sein, daß die ursprünglichen strukturellen Zusammenhänge, die Kontextinformationen und der Inhalt dem originalen Intranetauftritt entspricht⁵ und im Zuge der Archivierung nicht geändert wird

⁴ Mit dem Einsatz von Dokumenten-Management-Systemen kann die Charakteristik dahingehend verändert werden, daß das Intranet nur die Benutzeroberfläche darstellt, die Dokumente über ein Workflow gesteuert und verwaltet werden.

⁵ Vgl. Kapitel 7.2. „Characteristics of a record“ (ISO 15489-1: Information and documentation – Records management – Part 1: General, 2001, S. 7).

Nach dem ICA Guide bestehen elektronische Aufzeichnungen aus Inhalt, Kontext und Strukturanga-

bzw. manipulierbar ist. Technische Angaben über erforderliche Hardware- und Softwarekomponenten für die Lesbarkeit und Interpretation der Informationen müssen zusätzlich verfügbar sein.

Der Archivierungsablauf sollte auf Standards und nicht-proprietären Lösungen basieren. Zudem sollte das Archivierungskonzept so entwickelt werden, daß es leicht zu handhaben und für die Archivierung anderer bankrelevanter Websites anwendbar ist.

Vorgehensweise

Um den Archivierungszielen näherzukommen, fanden im Vorfeld Gespräche mit Verantwortlichen aus dem Archiv, der Intranetredaktion und mit mehreren IT-Fachleuten statt. Des weiteren orientierte ich mich an bereits bestehenden Projekten zur Archivierung von Websites bzw. von einzelnen Webseiten. Die Impulse zu dieser Thematik stammen jedoch überwiegend aus dem Bibliotheksbereich. Zu nennen ist hier besonders das PANDORA-Projekt (Preserving and Accessing Networked Documentary Resources of Australia) der National Library of Australia. Weitere Meilensteine sind: die Archivierung von Internetseiten durch „The Internet Archive“, die Archivierung des schwedischen Internets im Kulturarw3-Projekt, sowie die Gründung der Vereinigung EWA (European Web Archive) und der Vereinigung europäischer nationaler Bibliotheken NEDLIB (Networked European Deposit LIBrary) zur Entwicklung einer Infrastruktur für digitale Publikationen. Aus diesen Projekten der Zusammenarbeit gehen weitere Projekte hervor.⁶

Nur wenige Archive wie die National Archives of Australia, das Stadsarchief Antwerpen oder das Archiv der sozialen Demokratie in Bonn widmen sich der Archivierung von Websites.⁷

Schließlich fand in der Intranetredaktion eine Analyse des Mediums bzgl. des Publikationsprozesses, der Datenhaltung, Datenverwaltung sowie der Verantwortlichkeiten

ben, die zusammen die Geschäftstätigkeit beweisen. „A record is recorded information produced or received in the initiation, conduct or completion of an institutional or individual activity and that comprises content, context and structure to provide evidence of the activity.“ (Kapitel 2.1. in: International Council on Archives (ICA), Guide for managing electronic records from an archival perspective, hg. vom committee on Electronic Records, ICA Studies 8. Paris 1997)

<http://www.ica.org/biblio/cer/guide_12.html#top>.

⁶ PANDORA: <<http://pandora.nla.gov.au/index.html>>;

The Internet Archive: <<http://www.archiv.org>>;

Kulturarw3: <<http://www.kb.se/kw3/>>;

NEDLIB: <<http://www.kb.nl/coop/nedlib/>>.

⁷ National Archives of Australia: http://www.naa.gov.au/recordkeeping/er/web_records/intro.html;

Stadsarchief Antwerpen http://www.antwerpen.be/david/nl/text_websites.htm;

Archiv der sozialen Demokratie, Bonn: SCHMITZ, Rudolf: Archivierung von Intranetseiten.

Spiegelungsprojekt im Archiv der sozialen Demokratie (AdsD). In: Der Archivar 55 (2002), S. 135–136.

statt. Das Ergebnis war, daß die Verantwortlichkeiten als auch die Publikation sowohl zentral bei der Intranetredaktion als auch dezentral bei den einzelnen Unternehmensbereichen liegen. Die Idee, die Informationen über das zentrale CMS ins Archivsystem zu exportieren, wäre demnach nicht sinnvoll. Denn nur ein Teil der Intranet-Contents sind in dem CMS enthalten.

Allgemeines zur Archivierung von Websites

1. Herangehensweise zum Herunterladen der Web-Inhalte

Aus der Analyse bereits durchgeführter Projekte sind zwei Herangehensweisen zum Herunterladen der Web-Inhalte zu unterscheiden.

§ „comprehensive approach“

Nach der vollständigen Herangehensweise werden aus dem gesamten Web alle Webseiten aus einem bestimmten Webbereich vollständig mittels eines Web Crawler´s heruntergeladen. Diese Technik wird bereits bei Suchmaschinen wie Google oder Altavista angewendet. Die vollständige Vorgehensweise wird überwiegend von Nationalbibliotheken eingesetzt, deren Ziel es ist, einen länderspezifischen Teil des Internets zu sichern.

Das Ergebnis wird in einem Ranking dargestellt. Die Web-Contents stehen nicht im ursprünglichen Zusammenhang.

§ „selective approach“

Mit der selektiven Herangehensweise werden bereits ausgewählte Website mittels einer Spiegelungssoftware (Offline Browser) archiviert. Die Software lädt alle Dateien, ausgehend von einer bestimmten URL eventuell bis zu einer bestimmten Spiegelungsgrenze oder nach bestimmten Filterregeln herunter. Die selektive Vorgehensweise wird überwiegend von Archiven eingesetzt, deren Anzahl von Websites begrenzt ist. Die Ansicht des Spiegelungsergebnisses im Browser entspricht dem des ursprünglichen Webauftrittes.

2. Metadaten

Für die Verwaltung und Erhaltung authentischer webbasierter Aufzeichnungen nimmt die Erfassung von Metadaten eine zentrale Stellung ein.⁸

⁸ „A successful preservation process relies to a large extent on description of the nature and history of the archival resources: on metadata, in short. Metadata is also required to guide the way archived objects are rendered to and understood by users.“ (National Library of Australia: Archiving the Web: The PANDORA Archive at the National Library of Australia. Canberra 2001)
<<http://www.nla.gov.au/nla/staffpaper/2001/cathro3.html>>.

Metadaten zum Kontext und zur Struktur der Aufzeichnungen sind dabei notwendig, um die Aufzeichnungen zu verstehen und zu benutzen.⁹

Es existieren bereits Metadatenkonzepte, welche bei der Publikation und welche bei der Archivierung verwendet werden können. Für die Erschließung von Internetquellen gibt es den Bibliotheksstandard Dublin Core¹⁰, deren Metadata-Elemente in den Meta-Tags der Quellcodes Anwendung finden können. Die Meta-Tags im Quellcode der Intranetseiten sind jedoch unzureichend, fehlerhaft und oft nicht einheitlich. Deshalb müssen zusätzliche Metadaten mit technischen, administrativen Angaben, Angaben zum Spiegelungs- und Archivierungsprozeß verfügbar sein.

3. Lösung der technischen Frage

Für die Archivierung von Websites muß letztlich auch die Frage geklärt werden, welche Technologie notwendig ist, um die Informationen in einem authentischen Zusammenhang zugänglich zu machen und wiederherstellen. Peter Lyman faßt die Lösung der technischen Frage in drei Punkten zusammen:¹¹

- § die Überführung der Informationen auf eine neue Plattform (Migration)
- § die Überführung der Informationen in ein neues, standardisiertes Dateiformat
- § die Auswahl des Speichermediums

Archivierungsablauf

Aus den genannten Kriterien wurde folgender Archivierungsablauf entwickelt:

1. Intranet

Das Intranetportal ist der Einstieg zu verschiedenen Informationen, die sowohl innerhalb als auch außerhalb des Intranets (Intranet_x) sich befinden können.

2. Bewertung

Da nicht das Ziel ist, alle enthaltenen Informationen zu archivieren, muß eine Bewertung vorgenommen werden. Die Bewertungsentscheidungen sind zu protokollieren und ständig zu überprüfen.

⁹ „... This is an important concept for electronic records because metadata about the context and structure of a record is needed to make the record understandable and usable. As stated in the concept of a record, information about context is one of the necessary elements in providing evidence of the activity the record represents.“ (Kapitel 2.2, in: ICA, Guide, 1997).

¹⁰ Vgl. RUSCH-FEJA, Diann (Übers.): Metadata-Tags zur Erschließung von Internetquellen. Hg. von der Bibliothek und Wissenschaftlichen Dokumentation, Max-Planck-Institut für Bildungsforschung, Stand 18.12.1996. Berlin 1997 <<http://www.mpib-berlin.mpg.de/DOK/metatagd.htm>>.

¹¹ LYMAN, Peter: Archiving the World Wide Web. In: Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving, copublished by the Council on Library and Information Resources and the Library of Congress. Washington 2002 <www.clir.org/pubs/reports/pub106/contents.html>.

3. Spiegelung

Die Spiegelungssoftware HTTrack¹² wird zum Herunterladen der Dateien aus dem Intranet eingesetzt. Entsprechend den Bewertungskriterien werden alle Dateien und Ordner nach der originalen Datenstruktur hierarchisch in ein Verzeichnis auf dem Archivserver abgelegt. Als Ergebnis liegt ein Snapshot des Intranets auf dem Archivserver vor. HTTrack erstellt zu jeder Spiegelung eine Protokolldatei im ASCII-Format, die im gleichen Verzeichnis automatisch gespeichert wird.

4. XML-Dokument mit Metadaten

In einer XML-Vorlagedatei werden zusätzlicher Metadaten zum Snapshot erfaßt. Das XML-Dokument wird im gleichen Verzeichnis wie die Dateien des Snapshots abgelegt. Vorbereitende Dokumente können als Ergänzung für die Dokumentation dienen (z.B. Struktur der Website/Webseite, Installationsrichtlinien, Systembeschreibung, Richtlinien für den Administrator, Styleguide, o.ä.).

Die fehlenden Metadaten aus dem Quellcode können jedoch nicht rückwirkend ergänzt werden. Deshalb ist es wichtig, bereits bei der Erstellung der Webseiten aussagekräftige Metadaten anzugeben!

5. Migration

Liegt der Snapshot auf dem Archivserver vor, sollte im Anschluß eine Migration der Dateien in eine xml-basierte Sprache durchgeführt werden. Für die Migration von Dateien im HTML-Format in das XHTML-Format kann der HTML-XML-Konverter, der im Auftrag der Allianz Group entwickelt worden ist, an die Intranetdateien angepaßt werden. Drohen andere Text-, Grafik-, Audio- oder Videodateiformate oder Javascripte zu veralten, müssen schließlich auch für diese Formate Migrationsmethoden angewendet werden. Dafür gibt es aber noch keine. Das Ergebnis ist ein Snapshot des Intranets mit migrierten Dateien.

6. Indexierung des Snapshot

Aus der gespiegelten und migrierten Intranetseite sowie aus dem XML-Dokument mit Metadaten wird ein Index mittels einer Indexierungssoftware erstellt. Es besteht bereits eine Vielzahl von Softwareprodukten zur Indexierung des Webs. Erfahrungen zu dieser Thematik gibt es bereits aus dem Projekt der Nordic Web Archives.¹⁴

Der Index wird auf dem Server abgelegt.

¹² Siehe <<http://www.httrack.com>>.

¹³ Mit dem Projektbeginn im Historischen Archiv der Dresdner Bank im September 2003 haben sich die Prioritäten auf die Entwicklung eines Migrationstools für nicht-proprietäre Dateiformate (Word, Excel, PDF) in ein xml-basiertes Archivierungsformat verschoben.

¹⁴ Vgl. <<http://nwa.no/aboutNwaT.php>>.

7. Recherche und Zugriff über den Gesamtindex

Um auf einen bestimmten Snapshot oder eine bestimmte Intranetseite zugreifen zu können, muß eine Suche über die gesamten Index möglich sein.

8. Datenhaltung, Datensicherung auf separaten Speichermedium

Für die Indexierung und Benutzung werden die Daten auf einem Server mit ausreichender Speicherkapazität vorerst in der File-Struktur abgelegt. Für die Verwaltung der hierarchisch abgelegten Daten existiert bisher keine Lösung.

Die Datensicherung erfolgt über ein separates Speichermedium (z.B. CD-ROM oder DVD-R), das eine ausreichende Speicherkapazität für den Snapshot hat, eine lange Lebensdauer aufweist und für eine Benutzung geeignet erscheint.

Das Medium sollte regelmäßig kontrolliert und erneuert werden.

9. Benutzung

Die Benutzung der Snapshots aus dem Web-Archiv bzw. Intranet-Archiv erfolgen über entsprechende Startseiten, die noch gestaltet werden müssen. Genaue Vorgaben zur Präsentationsform existieren noch nicht.

Zu beachten ist auch, daß die Benutzer-PCs mit der entsprechenden Hardware und Software ausgestattet sein müssen, damit die Inhalte und Anwendungen lesbar und interpretierbar bleiben. Dazu zählen u.a. der Browser sowie Text-, Bild- und Videoverarbeitungsprogramme.

Die Bewertung

Die Bewertung findet im vorarchivischen Aufgabenfeld, während der Publikationsphase der Intranetseiten, statt. Im Archivierungsprozeß des Intranets steht die Bewertung am Anfang.

1. Bewertungskonzepte

Im Rahmen der Untersuchung wurde festgestellt, daß es für die Auswahl von Websites keine allgemeingültigen Richtlinien gibt. Vielmehr gibt es verschiedene Bewertungskonzepte nebeneinander, die an das Medium angepaßt werden müssen.

§ Allgemeine archivische Bewertungskriterien aus dem „Handbuch der Wirtschaftsarchive“:

Charakteristisch für das Intranet ist, daß es den Mitarbeitern firmeninterne und -externe Informationen bereitstellt. In der Papierwelt entspräche es einer umfassenden, regelmäßig erscheinenden Unternehmenspublikation, deren Ausgaben

vollständig vorliegen sollten.¹⁵ Denn die Publikation gibt Auskunft über Ziele und Schwerpunkte der Unternehmenstätigkeit, mit der sich häufig Zuständigkeiten und Organisation des Unternehmens ändern. Dazu zählen u.a. Beteiligungen, Tochtergesellschaften und Fusionen. Publikationen, öffentliche Verlautbarungen des Unternehmens, Geschäftsberichte, Image- und Produktbroschüren, Pressemitteilungen und interne Rundschreiben stellen eine wichtige Quelle für die spätere Forschung dar.¹⁶ Diese sind vollständig dauerhaft aufzubewahren.

Die Ausführungen von Renate Köhne-Lindenlaub im „Handbuch für Wirtschaftsarchive“ über archivwürdige Materialien sind mit denen des Historischen Archivs der Dresdner Bank identisch.¹⁷ Im Intranet befindet sich eine Vielzahl der genannten Publikationsarten. Da die Archivierung des gesamten Intranets nicht im Sinne des Historischen Archivs ist, müssen Bewertungsentscheidungen nach formalen, systematisch-inhaltlichen und nachfrageorientierten Kriterien getroffen werden. Sie werden für die Bewertung Ein Ziel der Bewertung ist, Doppelüberlieferungen zu vermeiden. Das betrifft zum Beispiel Online-Publikationen, die in gedruckter Form vorliegen. Auswahlkriterien dafür wurden im Rahmen des PANDORA-Projektes entwickelt.¹⁸ Weitere Kriterien sind der Aggregierungsgrad der dargestellten Informationen, die Lesbarkeit, die Interpretierbarkeit, die Zugänglichkeit, die Aufbewahrungsfristen (z.B. für aktuelle Geschäftsinformationen/Rundschreiben¹⁹) sowie die zu erwartende interne als auch externe Nutzernachfrage.²⁰

¹⁵ KÖHNE-LINDENLAUB, Renate: Erfassen, Bewerten, Übernehmen. In: KROKER, Evelyn u.a. (Hg.): Handbuch der Wirtschaftsarchive. Theorie und Praxis. München 1998, S. 118f.

¹⁶ EBD.

¹⁷ Archivwürdige Materialien sind für das Historische Archiv: Geschäftsberichte, Haus- und Kundenzeitschriften, Festschriften zu Jubiläen, Gebäude- und Zweigstelleneinweihungen und Fusionen, Dokumentationen, veröffentlichte Artikel, Vorträge, Bücher von Mitgliedern der Vorstände oder sonstiger Leitungsorgane, Presseverlautbarungen, und Werbematerialien.

¹⁸ Vgl. National Library of Australia: National Library of Australia: Guidelines for Selection of Online Australian Publications Intended for Preservation by the National Library of Australia. Canberra 2001 <<http://pandora.nla.gov.au/selectionguidelines.html>>.

¹⁹ Die gesetzliche Aufbewahrungsfrist für interne Rundschreiben als Arbeitsanweisungen oder Organisationsunterlagen, die zum Verständnis der Buchführung erforderlich ist, beträgt 10 Jahre (DAUEN, Sabine: Aufbewahrungspflichten. Von Originaldokumenten bis zur elektronischen Archivierung. Vorschriften, Fristen, Nachweispflichten, Vernichtung. Freiburg i.Br. 2002, S. 92). Auch die Dresdner Bank besitzt einen internen Fristenkatalog. Inwiefern die Aufbewahrungsfristen für Intranetangebote durch die verantwortlichen Unternehmensbereiche eingehalten werden, kann nicht kontrolliert werden. Da Rundschreiben und Geschäftsberichte für das Archiv sowieso archivwürdig sind, hat dieses Kriterium keine Auswirkung auf die Auswahl der Intranetangebote.

²⁰ KÖHNE-LINDENLAUB, Bewertung, S. 109.

§ Allgemeine Bewertungsrichtlinien für digitale Aufzeichnungen:

Allgemeine Richtlinien für die archivische Bewertung digitaler Informationen sind in *Preserving Digital Information*²¹ aufgeführt. Es ist das Ergebnis der Arbeitsgruppe *Task Force on Archiving of Digital Information* und bildet die Grundlage für verschiedene Projekte, die sich mit der Archivierung von einzelnen Websites befassen.²² Darin heißt es, daß Auswahlkriterien eine Bewertung nach dem Inhalt des Objektes in Beziehung zum Sammlungsziel des digitalen Archivs, der Qualität und der Einzigartigkeit des Objektes beinhalten.²³

Für die Bewertung von Webseiten müssen die allgemeinen Richtlinien in *Preserving Digital Information* für Hypertextdokumente angepaßt bzw. umgesetzt werden.

- Ausgehend vom Ziel des Projektes, werden erste Bewertungsentscheidungen getroffen, die von der Herangehensweise abhängen. Projekte der vollständigen Herangehensweise definieren die Top-Level-Domain,²⁴ um diese mittels „Web Crawler“ zu archivieren. Sie führen eine Bewertung auf Makroebene durch.
- Projekte der auswählenden Herangehensweise wählen z.B. elektronische Publikationen, Diskussionslisten, einzelne Websites oder nur eine Website aus. Für die auswählende Archivierungsmethode gibt es Richtlinien zur Auswahl von Online-Publikationen und für die Einschätzung des (Archivierungs)Risikos, die als Grundlage für die Spiegelungstiefe, Archivierungsintervalle und den Einsatz bestimmter Filter dienen.

§ Auswahlkriterien für Online-Publikationen:

In *Guidelines for the Selection of Online Australian Publications Intended for Preservation by the National Library of Australia*²⁵ werden Richtlinien zur Auswahl von Online-Publikationen gegeben. Die Nationalbibliothek verfolgt das Ziel, nur australische Online-Publikationen auf Dauer zugänglich zu machen. Dazu zählen die traditionellen Sammlungsobjekte von Bibliotheken (Bücher, Magazine, Zeit-

²¹ Task Force on Archiving of Digital Information (Hg.): *Preserving digital Information*, Report of the Task Force on Archiving of Digital Information, commissioned by The Commission on Preservation and Access and The Research Libraries Group, 1996 <<http://www.rlg.org/ArchTF/>>.

²² Vgl. LYMAN, *Archiving WWW*, 2002.

²³ „In general, selection criteria include an appraisal of the content of the object – its subject and discipline – in relation to the collection goals of the digital archives, the quality and uniqueness of the object.“ (Task Force, *Preserving Information*, 1996, S. 21).

²⁴ Zum Beispiel: „The Internet Archive“ – .com; „Kulturarw3-Projekt“ – .se

²⁵ National Library of Australia: *Guidelines for Selection of Online Australian Publications Intended for Preservation by the National Library of Australia*. Canberra 2001 <<http://pandora.nla.gov.au/selectionguidelines.html>>.

schriften, Zeitungen u.ä.), die öffentlich über das Internet in elektronischer Form zugänglich sind. Organisatorische Aufzeichnungen oder einzelne Materialien, die im Zuständigkeitsbereich von Archiven liegen, werden von den National Archives of Australia bewertet und archiviert.

- Die National Library of Australia erklärt die Notwendigkeit von Auswahlkriterien damit, daß nicht alle Versionen bzw. Editionen aufbewahrt werden können. Die Festlegung der Archivierungsintervalle hängt ab von dem Publikationsmuster, der Bedeutung enthaltener Informationen sowie der Stabilität der Website. So sollten einige „Titel“ so vollständig wie möglich vorliegen. Bei anderen dagegen würden einige Snapshots ausreichen.
- Liegt die australische Online-Publikation in anderen Speichermedien vor, wie z.B. in Papierform oder auf Mikrofilm, wird die Online-Version nur aufbewahrt, wenn sie wichtige zusätzliche Informationen oder Bedeutung hat. Sonst wird generell die gedruckte Form bevorzugt. Sind die gleichen Informationen sowohl online als auch auf CD-ROM oder auf Diskette verfügbar, wird die Online-Version bevorzugt.
- Eine hohe Priorität haben maßgebliche Publikationen mit einem hohen zu erwartenden Nachfragewert. In der weiteren Ausführung zu „authoritative publication“ wird erklärt, daß zu einem authentischen Titel sowohl die Organisation bzw. Person, die für die Publikation verantwortlich ist, als auch die Qualifikation des Webpublishers bekannt sein sollte.
- Entspricht die Online-Publikation nicht den genannten Kriterien zur Authentizität und Nachfrage, wird sie auch nicht archiviert.
- Publikationen mit einem geringem Aggregierungsgrad, deren Informationen bereits woanders vorliegen, werden nicht aufbewahrt.
- Informationen über Sponsoren oder sonstige Unterstützungen können die Bewertung zur Auswahl positiv beeinflussen.
- Auch innovative Beispiele von Internetpublikationen, die für die spätere Forschung interessant sein könnten, werden für die dauerhafte Aufbewahrung ausgewählt.
- Die Auswahl nach thematischen Gesichtspunkten dient nur dazu, die gedruckte Sammlung zu ergänzen bzw. mit weiterreichenden Informationen zu vervollständigen. Für Australien werden die Beispiele der Aborigines und der Olympischen Spiele genannt.

- Wichtig für die Auswahl von Websites ist es, die Grenzen zu definieren. So sollten nur interne Links archiviert werden.
- Größere Websites sollten in einzelne Komponenten zerlegt und diese dann nach den einzelnen Bewertungskriterien ausgewählt werden. Können die Komponenten dagegen nicht separat betrachtet werden, sollte die Website als Ganzes archiviert werden.
- Neben den allgemeinen Auswahlrichtlinien gibt es auch Regeln für spezielle Publikationsformen, wie Jahresberichte, digitalisierte Materialien oder aber auch Zeitungen.
- Für online-verfügbare Jahresberichte gilt, daß diese nur dann ausgewählt werden, wenn sie nicht in gedruckter Form vorliegen oder nicht regelmäßig in anderen Publikationsformen (wie CD-ROM oder Diskette) erscheinen. Bei Zeitungen und Zeitschriften wird ähnlich wie bei Jahresberichten vorgegangen. Online-Publikationen werden nicht aufbewahrt, wenn die Informationen der gedruckten Ausgabe nur dupliziert werden.
- Digitalisierte Materialien, deren Original z.B. in Papierform vorliegt, werden generell nicht aufbewahrt, da die National Library of Australia davon ausgeht, daß die Abteilungen ihre eigenen digitalen Materialien sichern. Es gibt aber auch Ausnahmefälle, die die Bewertungsentscheidung beeinflussen. Demnach werden digitalisierte Materialien ausgewählt, wenn es sich erstens um historische Dokumente handelt und deren Archivierung unterstützt werden soll und zweitens, wenn die Webseite mehr als eine digitalisierte Kopie enthält.
- Im Unterschied zum archivischen Aufgabenfeld, daß von den National Archives of Australia bearbeitet wird, wählt die National Library of Australia keine organisatorischen und persönlichen Seiten aus. Gleiches gilt für Entwürfe und Arbeiten, die noch im Prozeß sind. Denn es sollen möglichst nur vollständige, komplette Dokumente ausgewählt werden.

§ Richtlinien zur Risikoeinschätzung von Websites:

Für die Bewertung von einzelnen Webauftritten ist nach dem *Guidelines for Keeping Records of Web-based Activity in the Commonwealth* und der Publikation *An Approach to Managing Internet and Intranet Information for Long Term Access and Accountability* eine Risikoanalyse der Webseiten entscheidend. Wie oft ein

Snapshot einer Website durchgeführt werden soll, hängt von dem Ergebnis der Einschätzung ab.²⁶

Bei der Risikoanalyse wird mit Blick auf die dauerhafte Aufbewahrung des Snapshots untersucht, welche Probleme auftreten, warum und was dagegen getan werden kann. Das Risiko bezieht sich auf den Verlust der Daten und dem Verwaltungsaufwand, um dies zu vermeiden. Man spricht auch von „risk management“.²⁷

Wird das Risiko hoch eingeschätzt, müssen mehr Anstrengungen für die Bewahrung und Erhaltung der digitalen Aufzeichnungen unternommen werden. Dies beeinflusst die Festlegung der Archivierungsintervalle und die Auswahl bestimmter Websites oder einzelner Webseiten.

2. Risikoeinschätzung der National Archives of Australia

Die Richtlinien zur Einschätzung des Risikos durch die Abteilungen der Commonwealth-Regierung werden ausführlich in *Guidelines for Keeping Records of Web-based Activity in the Commonwealth* beschrieben. Darin heißt es, daß die Bewertung das Ergebnis einer Risikoanalyse ist. Sie wird von der entsprechenden Abteilung, die für die Inhalte der Webseite verantwortlich ist, durchgeführt.

Wird ein hohes Risiko ermittelt, müssen die Aufzeichnungen für eine dauerhafte Aufbewahrung vollständig erfaßt werden. Im Hintergrund liegende Informationen müssen freigelegt und verfügbar sein (z.B. Informationen aus Datenbanken).

Für die Einschätzung des Risikos werden von den National Archives of Australia vier Faktoren genannt:²⁸

- Sichtweise der Öffentlichkeit auf die Behörde
- Ziel der Website
- Komplexität der Website
- Häufigkeit und Regelmäßigkeit der inhaltlichen Änderungen

²⁶ National Archives of Australia: Archiving Web Resources: Guidelines for Keeping Records of Web-based Activity in the Commonwealth Government. Canberra 2001

<http://www.naa.gov.au/recordkeeping/er/web_records/intro.html>;

IM (Information Management) Forum Internet and Intranet Working Group: An Approach to Managing Internet and Intranet Information for Long Term Access and Accountability, 24.

September 1999 <http://www.imforumgi.gc.ca/iapproach2_e.html>.

²⁷ „Risk management is defined as the systematic application of policies, procedures and practices to the tasks of identifying, analysing, evaluating, and monitoring risk. Essentially it involves anticipating what can go wrong, why and what can be done.“ (Appendix 11 – Risk analysis in DIRKS, in: National Archives of Australia: Risk analysis in DIRKS, in: DIRKS Manual, 2003)

<http://www.naa.gov.au/recordkeeping/dirks/dirksman/dirks_A11_risk.html>.

²⁸ Vgl. NAA, Guidelines, 2001, S. 20–23.

3. Risikoeinschätzung des Information Management Forums

Die Einteilung der Risikoeinschätzung in *An Approach to Managing Internet and Intranet Information for Long Term Access and Accountability* des IM Forums basiert auf dem Model der Autoren Charles R. McClure und J. Timothy Sprehe.²⁹

§ Ein *geringes Risiko* liegt vor, wenn die Inhalte einer Website unter einer wirkungsvollen Kontrolle publiziert werden. Dies ist auch der Fall, wenn die Website überwiegend aus statischen Seiten und wenigen interaktiven Element besteht, die Inhalte alle im Datenhaltungssystem vorliegen (z.B. CMS) und enthaltene Online-Publikationen noch in anderen Speichermedien publiziert

§ Ein *mittleres Risiko* besteht bei Websites, deren Kontrolle nicht ausreichend bei einer wachsenden Anzahl von Webseiten ist. Die Webseiten sind teils statisch teils dynamisch, mit einer zunehmenden interaktiven Funktion. Die Contents liegen nicht alle im Datenhaltungssystem vor, und für Online-Publikationen existiert nicht immer eine äquivalente Printausgabe.

§ Bei Websites mit einem *hohen Risiko* spitzt sich die Situation zu. Eine Kontrolle über die Inhalte ist nur sehr schwer möglich. Die Webseiten sind dynamischen und interaktiven Charakters, liegen überwiegend nicht im Datenhaltungssystem vor, und viele Online-Publikationen haben keine gedruckte Ausgabe mehr.

4. Bewertungsstrategie

Für die Umsetzung der Bewertungskriterien werden von den National Archives of Australia zwei Bewertungsstrategien³⁰, „object-driven“ (objektgesteuert) oder „event-driven“ (ereignisgesteuert), vorgeschlagen. Für welche der beiden sich entschieden werden sollte, hängt ab von der Komplexität der Webquelle, der Art der webbasierten Interaktionen, von dem Ergebnis der Risikoeinschätzung sowie der Analyse für die Anforderungen an die Datenhaltung. Eine Kombination der beiden Strategien wird jedoch empfohlen.³¹

Die objektgesteuerte Bewertungsstrategie ist geeignet für statische Websites, die eine Sammlung von HTML-Dokumenten umfassen und keine komplexen Interaktionen enthalten. Folglich könnten Snapshots in periodischen Zeitabständen von der Website gemacht werden.

²⁹ Vgl. McCLURE, C. R./SPREHE, J. T.: Analysis and Development of Model Quality Guidelines for Electronic Records Management on State and Federal Websites, 1998 (zit. aus: IM Forum, Approach, 1999)

³⁰ Vgl. NAA, Guidelines, 2001, S. 24.

³¹ „Agencies may consider object-driven or event-driven approaches or, better still, a combination of the two“ (EBD.).

Die ereignisgesteuerte Bewertungsstrategie wird empfohlen für Webauftritte, die aus einer Reihe von interaktiven oder dynamisch generierten Webseiten bestehen. Sie liefern eine Antwort unikaten Charakters auf eine bestimmte Anfrage. Diese Vorgehensweise ist Bestandteil einer funktionalen Bewertung für die Geschäftstätigkeit einer Abteilung. Denn damit werden auch Ereignisse oder Transaktionen erfaßt, die zwischen der Webseite und dem Nutzer stattfinden.³²

Bewertung des Intranets

1. Allgemeines

Die Gesamtheit der aufgeführten Bewertungskriterien können nicht 1:1 für die Bewertung des Intranets übertragen werden. Die Grundprinzipien der Bewertung, dargestellt im „Handbuch für Wirtschaftarchive“, müssen an das digitale Medium angepaßt werden. Die Risikoanalyse, wie sie von den National Archives of Australia und des IM Forums vorgeschlagen wird, hat für die Bewertung des Intranets nur für das Kriterium der „Zugänglichkeit und Erhaltungszustand“ der Aufzeichnungen eine Bedeutung. Aus diesem Grund wird auf eine ausführliche Einschätzung des Risikos für das Intranet verzichtet.

2. Bewertungskriterien

Aus den verschiedenen Bewertungsansätzen kristallisierten sich folgende Bewertungskriterien heraus:

- Ziel der Archivierung
- archivischer Zuständigkeitsbereich des Historischen Archivs der Dresdner Bank
- Gesamtfunktion der Abteilung bzw. des Intranetangebotes im Unternehmen
- Evidenzwert für die Dresdner Bank
- Aggregierungsgrad der enthaltenen Informationen
- Häufigkeit und Regelmäßigkeit der Änderungen
- Redundanzen vermeiden
- Recherche- und Nachfragewert
- Zugänglichkeit und Erhaltungszustand
- Benutzung

³² Dazu gehört: Benutzerprofil, Style Sheets, Suchmaschinen, Skripte und Programme, regelmäßige Snapshots der Datenbank selbst und Datenbanktransaktionsprotokolle („database transaction logs“).

3. Umsetzung der Kriterien

Die Software bietet die Möglichkeit, die entwickelten Bewertungskriterien sehr präzise umzusetzen. Unterschiedliche Archivierungsintervalle mit verschiedenen Einstellungen können schnell und einfach durchgeführt werden. Mehrere Spiegelungsgrenzen (z.B. interne und externe Verzeichnungstiefen, Linkanzahl, Anzahl der Verbindungen b/s oder maximale Speicherkapazität des gesamten Snapshots) können gesetzt werden. Zudem bietet die Software umfangreiche Funktionen zum Einschluß bzw. Ausschluß von Links; Datei- und Pfadnamen oder Server- und Domänebezeichnungen sind vorhanden und können einfach bedient werden. Die Spiegelungstests zeigten aber auch, daß Bewertungsentscheidungen nach Arbeitsaufwand und Aspekt der Speicherkapazität getroffen werden.

§ Der Arbeitsaufwand für die Bewertung einzelner Webseiten (bei mehr als 800.000) ist viel zu hoch. Vergleicht man Bewertungsentscheidungen in der Papierwelt, wird auch dort im allgemeinen auf eine Einzelblattkassation verzichtet.

§ Die technische Entwicklung beeinflußt auch die Kosten für den Speicherplatz und die Speicherkapazität neuer Speichermedien. Das heißt, Speicherplatz wird immer günstiger und die Speichermedien umfassen immer mehr Speicherkapazität.

Diese Argumente werden den präzisen Einstellungsmöglichkeiten der Software entgegen gesetzt, damit der Archivierungsprozeß leicht handhabbar bleibt und automatisch ablaufen kann.

Zusammengefaßt wird die Bewertung von Webseiten in Form der Spiegelungsgrenzen, Spiegelungsintervalle sowie durch (allgemeine) Einschluß- und Ausschlußfilter wie z.B. `+*.dresdner.net/*` `-*allianz*`, `-www.*` oder `-*telefonbuch*`

Sicherung und Erhaltung des Intranet-Archivs

1. Allgemeines

Die Snapshots des Intranets liegen nach der Spiegelung auf der Festplatte im Archiv vor. Sie sollen nun auf Dauer gesichert und erhalten werden. Die kurze Lebensdauer der physikalischen Speichermedien und die schnelle Entwicklung von Hard- und Software erfordern weitere Archivierungsstrategien. Sie ist eng verbunden mit dem Archivierungsziel – warum die Snapshots aufbewahrt werden sollten.³³ Es wird dabei zwischen drei Schwerpunkten unterschieden:

³³ ARMS, William Y. u.a.: Collecting and Preserving the Web: The Minerva Prototype. In: RLG DigiNews (2001), Vol. 5., No. 2 <<http://www.rlg.org/preserv/diginews/diginews5-2.html#feature1>>.

- § „Preservation of bits“ - Ziel ist es, die exakte Bitsequenz zu erhalten, wie sie im Original vorliegt
- § „Preservation of content“ – Ziel ist es, den Inhalt jedoch nicht die volle interaktive Natur der Website zu erhalten (z.B. Text, Grafiken, Audio)
- § „Preservation of experience“ – Ziel ist es, das vollständige Erlebnis der Interaktionen mit dem digitalen Material zu bewahren, einschließlich dem Sehen und Fühlen sowie den Ausführungen von dynamischen Elementen.

Da das Archivierungsziel neben der Erhaltung der Inhalte auch die Erhaltung der Charakteristik des Mediums, also der Recherchierbarkeit und der Navigation, ist, wird die Funktionalität des Snapshots durch folgende Komponenten beeinflusst.³⁴

- Verwendung von unterschiedlichen Versionen und Typen von HTML, die auch verschiedene Funktionalitäten besitzen
- Plattformabhängigkeit von Software, Suchmaschinen oder datenbankbasierten Fragetools
- Korrekte Einbettung und Verlinkung von verschiedenen Versionen von Anwendungen, die für deren Funktionalitäten inklusive Applets, JavaScript und Software plug-ins gebraucht werden
- Beschränkungen von einigen (älteren) Browsern
- Geschätzte physikalische und/oder kommerzielle Lebensdauer des Mediums, auf dem die Snapshots und die Metadaten gespeichert sind
- Langzeitverfügbarkeit der Hardware und der Plattform des Betriebssystems, die notwendig ist, um den Zugang zu den gespeicherten Aufzeichnungen auf verschiedenen Medientypen zu erhalten.

Die dauerhafte Erhaltung webbasierter Aufzeichnungen umfaßt demnach Aufgaben:³⁵

- Auswahl der Archivierungsstrategien; das regelmäßige Überprüfen und Erneuern der Speichermedien sowie die Migration
- Verwendung von weit verbreiteten Standards, nicht allein für die Datenformate, sondern auch für Programme und Softwaretools; Umsetzung nicht-proprietärer Lösungen
- Einführung von Sicherheitsmerkmalen gegen bewußte oder unbewußte Änderungen
- Verwendung von dauerhaften Identifikationen/Signaturen
- Sicherung von Kontroll- und Überwachungsmechanismen über die Umwelteinflüsse

³⁴ Vgl. NAA, Guidelines, 2001, S. 32.

³⁵ Vgl. EBD., S. 33f.

- Auswahl des Speichermediums

2. Dateiformate für die Archivierung

Damit bei einer Migration keine Informationen verloren gehen und das Dokument authentisch bleibt, muß ein standardisiertes Speicherformat ausgewählt werden.³⁶ Das gilt nicht nur für den Archivierungsprozeß, sondern sollte schon im Vorfeld, bei der Erstellung der Webseiten, Anwendung finden.

Für die dauerhafte Bewahrung der webbasierten Aufzeichnungen gelten im Prinzip die gleichen Grundsätze wie für andere digitale Aufzeichnungen. Nach Mannerheim ist die Situation jedoch etwas einfacher, da 95% der Dateien in standardisierten Formaten vorliegen.³⁷

Für die Archivierung von Hyperlink-Dokumenten wird SGML und zu deren Präsentation HTML empfohlen. Mit Blick auf die Weiterentwicklung steht aus heutiger Sicht die Kombination beider Standards in XHTML zur Verfügung.³⁸

Bei der Auswahl von Grafik-, Audio- und Videodateiformaten sind, sowohl für die Archivierung als auch für das Publizieren im Intranet, neben der Verfügbarkeit, Konvertierbarkeit, Standardisierung, Recherchierbarkeit und Archivierbarkeit auch die Dateigröße, Qualität und Authentizitätsmerkmale der Datei entscheidend.³⁹ Auf diese Thematik wurde in der Diplomarbeit nicht weiter eingegangen.

3. Archivierungsstrategien

Die Lesbarkeit und die Zugänglichkeit der Informationen hängt nicht zuletzt von Archivierungsstrategien wie *Emulation* oder *Migration* ab. In allen Archivierungsprojekten für Websites wird der Weg der Migration gewählt, wobei die Vorgehensweise auch kritisch betrachtet wird.⁴⁰ Die Migration ist ein ständig fortlaufender Prozeß, der

³⁶ OHST, Daniel: Dateiformate für das elektronische Publizieren (Studienarbeit, Humboldt-Universität zu Berlin, Institut für Informatik). Berlin 1998 <<http://edoc.hu-berlin.de/buecher/ohst-daniel/HTML/>>. Der Schwerpunkt der Studienarbeit liegt auf Dateiformaten für das elektronische Publizieren.

³⁷ „Long-term preservation of web publications is in principle not different from long-term preservation of any other digital information. Maybe the situation is a little easier because over 95 percent of the files, HTML and image files, are in standardised formats. So the prospect of having software reading them in the future is better than in other areas using proprietary software.“ (MANNERHEIM, Johan: The WWW and our digital heritage – the new preservation tasks of the library community, 66th IFLA council and Conference, Conference Proceedings, Jerusalem 13–18 August 2000 <<http://www.ifla.org/IV/ifla66/papers/158-157e.htm>>).

³⁸ Vgl. <<http://www.w3.org/TR/xhtml1>>.

³⁹ Vgl. BÜTTNER, Stephan: Digitale Bibliothek. Elektronisches Publizieren: Formate, Vorlesungsmaterialien an der FH Potsdam. Potsdam SS 2002.

⁴⁰ Bei der Emulation werden Programme (Emulatoren) entwickelt, die auf zukünftigen Systemen, das Verhalten veralteter Betriebssysteme nachahmen. Mit der Migration werden die Daten entweder von einer Hardware/Software-Konfiguration auf die nächste oder aber sie werden von einer Generation der Computertechnologie auf die darauffolgende übertragen. (Nach: Task Force, Preserving Infor-

mehr Personal erfordert und hohe, permanente Kosten verursacht. Ein schwerwiegender Nachteil der Migration ist, daß dabei oft Informationen verlorengehen.⁴¹ Die Migration stellt aber für alle bisherigen Projekte die einzige umsetzbare Archivierungsstrategie dar. Zwar gibt es noch das „Refreshing“, bei dem die Informationen auf ein neues Medium kopiert werden. Die Informationen sind aber nur so lange lesbar, bis die Hard- und Software nicht mehr zur Verfügung steht. „Refreshing“ kann nur eine Lösung für kurze Zeit sein.⁴²

§ Migration von webbasierten Aufzeichnungen:

Wenn der Datentyp veraltet oder nicht für die Archivierung geeignet erscheint, muß es auf einen neuen, equivalenten Datentyp konvertiert werden. Die Migration muß bei der hohen Anzahl von Dateien, die zu einem gespiegelten Intranetauftritt gehören, automatisch für jede einzelne Datei durchgeführt werden. Um eine vollständige Webseite zu konvertieren, muß nicht nur die Dokumentenbeschreibungssprache HTML, sondern auch alle eingebunden Text-, Grafik- und Videodateien sowie Style Sheets und Skriptsprachen migriert werden.

Zum Beispiel: HTML 3.2→HTML 4.0→XHTML 1.0, CSS 1.0→CSS 2.0→XSL 1.0, Java 1.0→Java 2.0.

Theoretisch wird in den meisten Fällen der Inhalt konvertiert. Es kann aber auch die Darstellung bzw. das Ergebnis konvertiert werden.⁴³

Die Migration der Dateien sollte so oft wie nötig durchgeführt werden.⁴⁴ Dabei ist zu achten, daß die originalen Dateien immer erhalten bleiben, da eine Migration häufig mit Informationsverlusten verbunden ist.

Zum Ablauf der Migration ist ein Protokoll anzufertigen und mit in dem Verzeichnis des durchgeführten Spiegelungsprojektes abzulegen ist.

§ HTML-XML-Konverter:

Die Umwandlung von HTML 4.01 zu XHTML 1.0 ist theoretisch einfach durchzuführen, da XHTML bekanntlich eine Umformulierung von HTML 4.01 ist. Um Prob-

mation, 1996, S. 4f.), vgl. National Library of Australia: Archiving the Web: The PANDORA Archive at the National Library of Australia. Canberra 2001

<<http://www.nla.gov.au/nla/staffpaper/2001/cathro3.html>>

⁴¹ „...it is not always possible to make an exact digital copy or replica of a database or other informational object as hardware and software change and still maintain the compatibility of the object with the new generation of technology.“ (Task Force, Preserving Information, 1996, S. 5).

⁴² „Refreshing thus cannot serve as a general solution for preserving digital information.“ (EBD., S. 4)

⁴³ Vgl. ARMS, Minerva, 2001.

⁴⁴ „Web-based records and their associated metadata should be migrated as often as necessary to avoid technological obsolescence for as long as the records are required“ (NAA, Guidelines, 2001, S. 33).

leme bei der Validierung zu XHTML zu vermeiden, müssen die HTML-Dokumente gefiltert und korrigiert werden, da sonst viele Einbettungsfehler und Probleme bei XHTML auftreten können. Für die eigentliche Umwandlung existieren bereits automatische Konverter-Tools, die über das Internet aufrufbar sind. Sie enthalten Tidy, eine Anwendung, die HTML-Codes überprüft und korrigiert wird.⁴⁵ Darüber können jedoch nur einzelne Webseiten korrigiert und transformiert werden. Auf Grund der hohen Anzahl von HTML-Dateien, die nach einer Spiegelung vorliegen, ist ein Konverter erforderlich, der automatisch mehrere HTML-Dateien korrigiert und in XHTML umwandelt.

Im Auftrag der Allianz, entwickelte Herr Oehler, ein Mitarbeiter der Dresdner Bank, einen HTML-XML-Konverter, der diese Anforderungen erfüllt. Damit ist es möglich, ein Verzeichnis mit mehreren unkorrekten HTML-Dateien in reine XML-Dateien umzuwandeln. Dieser Konverter könnte für die Migration von HTML-Dokumenten nach XHTML angepaßt werden.

4. Fazit:

Abschließend ist zu festzustellen, daß die Archivierung von Websites bereits praktiziert wird und in der Umsetzung weniger Aufwand bedarf als erwartet. Es wird aber auch deutlich, daß sich an die Archivierung von Websites bisher nur wenige Archive gewagt haben. Aber enthält nicht auch der Internet- bzw. Intranetauftritt einer Behörde, Organisation, Institution oder eines Unternehmens wichtige Informationen, die für die zukünftigen Nutzer von Interesse sein könnten. Spiegeln sie nicht einen wichtigen Teil unserer heutigen Informationsgesellschaft wieder?

Die Diplomarbeit zeigt, daß es sich lohnt ein neues Gebiet der Archivierung zu betreten. Archivare sollten nicht davor zurückschrecken neue Informationsformen zu archivieren, sich mit der Technik auseinander zusetzen und über Archiv- und Ländergrenzen hinweg zu kooperieren.

⁴⁵ Tidy <<http://www.w3.org/People/Raggett/tidy/>>.