

Wilfried Enderle

## Kurzbericht von der Archiving Web Resources International Conference der National Library of Australia (9.-11. November 2004)<sup>1</sup>



Die australische Nationalbibliothek hatte als eine der ersten großen Bibliotheken begonnen, sich mit Fragen der Archivierung von Websites als Teil ihres nationalen kulturellen Erbes zu beschäftigen. Es lag daher nahe, daß die National Library of Australia zum Thema der Langzeitarchivierung digitaler Medien eine große, internationale Konferenz organisierte, auf der sowohl eine umfassende Bilanz bisheriger Aktivitäten als auch ein Ausblick auf künftige Konzepte gegeben wurde. Die Konferenz richtete sich im wesentlichen an Bibliothekare, aber auch an Archivare und Wissenschaftler, die sich mit der Erarbeitung von organisatorischen wie technischen Konzepten zur Archivierung digitaler Medienformen beschäftigen. Auf der Konferenz (ca. 200 Teilnehmer) waren im wesentlichen Vertreter der großen westlichen Nationalbibliotheken vertreten, der amerikanischen Bibliotheksverbände (OCLC, RLG), einiger amerikanischer Forschungsuniversitäten (Cornell, Harvard), Vertreter von Nationalarchiven (Großbritannien, Kanada, Australien) sowie weitgehend Bibliothekare aller wissenschaftlichen Bibliotheken Australiens und Neuseelands. Dazu kamen bibliothekarische Vertreter aus Japan und Singapur.

Die Konferenz selbst bot einen guten und vollständigen Überblick über aktuelle Archivierungsaktivitäten der wichtigsten Nationalbibliotheken sowie ausgewählter Nationalarchive, wobei sich die Vorträge im wesentlichen auf allgemeine konzeptionelle Themen beschränkten und weniger die komplexen technischen Detailfragen behandelten. Im folgenden soll nicht das gesamte Vortragsprogramm im einzelnen referiert, sondern einige ausgewählte Themen angesprochen werden, die für die konzeptionelle Planung größerer Forschungsbibliotheken von Interesse sein können.

Die über drei Tage dauernde Konferenz begann am ersten Tag mit vorbereitenden und hier nicht näher ausgeführten Grundsatzreferaten, deren Themenspektrum von der kulturellen und historischen Bedeutung des Web, den wissenschafts- und medienpolitischen Konsequenzen (Themen waren zum Beispiel Open Access oder die Zeitschriftenkrise) bis hin zur künstlerischen Nutzung des Web reichte. Am zweiten Konferenztag wurden dann konkrete Archivierungsaktivitäten vorgestellt. Dabei wurden die derzeit existierenden verschiedenen konzeptionellen Ansätze vorgestellt:

- Der Archivierung des gesamten Web (globaler Ansatz).
- Der Archivierung aller Websites eines Landes (vollständiger nationalbibliothekarischer Ansatz).
- Der Archivierung ausgewählter Websites eines Landes (selektiver nationalbibliothekarischer Ansatz).

- Der Archivierung ausgewählter Websites zu bestimmten Themenprofilen (fachbibliothekarischer Ansatz).

Der Archivierung des gesamten Web hat sich das in San Francisco beheimatete Internet Archive verschrieben, das seit 1996 regelmäßig (alle zwei Monate) einen snapshot des Web erfasst und zur Zeit pro Monat ca. 30 Terabyte Daten einsammelt<sup>2</sup>. Auch wenn es sich hier derzeit nur um ein reines robotgesteuertes Einsammeln und Speichern von Daten handelt, nicht um ein langfristiges Archivieren, so beginnt das Internet Archive mittlerweile doch enger mit bibliothekarischen Partnern zu kooperieren, denen es um die längerfristige Sicherung von Websites geht. So gibt es gemeinsame Projekte mit der Cornell University Library oder der Library of Congress (siehe unten), aber auch eine Kooperation mit dem International Internet Preservation Consortium (siehe unten).

Das Sammeln und Speichern aller Websites eines Landes hat in den letzten Jahren kontinuierlich die schwedische Nationalbibliothek betrieben (im Projekt Kulturarw<sup>3</sup>), wobei der freie Zugriff auf diese Ressourcen über das Internet derzeit aus urheberrechtlichen Gründen nicht möglich ist. Ergänzend zu dem Harvesten aller schwedischen Sites werden darüber hinaus ausgewählte digitale Bücher auch eigens katalogisiert und archiviert<sup>3</sup>.

Die australische Nationalbibliothek hat seit 1997 in dem Projekt Pandora begonnen, ausgewählte australische Websites zu archivieren<sup>4</sup>. Konzeptionell wie technisch handelt es sich dabei um eines der am weitesten entwickelten bibliothekarischen Archivierungsprojekte. Technisch deshalb, weil das von der National Library of Australia entwickelte System Pandas den gesamten Workflow unterstützt (siehe unten) und die Daten in einer Form speichert, die die Voraussetzung für die Langzeitarchivierung bieten; konzeptionell, weil sie von allen Sites die Genehmigung der Rechteinhaber einholen und damit die Möglichkeit besitzen, den Zugriff auf die aktuellen sowie ältere archivierte Sites auch frei über das Internet anbieten zu

<sup>1</sup> Zum Konferenzprogramm siehe <<http://www.nla.gov.au/webarchiving/index.html>> (Stand: 2.2.2005); vgl. auch den Bericht von Margaret E. Phillips: Archiving Web Resources International Conference: Issues for Cultural Heritage Organisations. In: RLG DigiNews Feb. 15, 2005 <[http://www.rlg.org/en/page.php?Page\\_ID=20522&Printable=1&Article\\_ID=1704](http://www.rlg.org/en/page.php?Page_ID=20522&Printable=1&Article_ID=1704)>.

<sup>2</sup> Internet Archive <<http://www.archive.org/>>.

<sup>3</sup> Kulturarw3-Long time preservation of electronic documents <<http://www.kb.se/kw3/ENG/Default.htm>>.

<sup>4</sup> Pandora <<http://pandora.nla.gov.au/index.html>>.

können. Derzeit hat die National Library of Australia ca. 7 000 Websites archiviert (ca. 700 GB; ca. 22 Mio. Dateien). Das zuständige Projektteam besteht dafür aus sieben Mitarbeitern. Das Konzept und voraussichtlich auch das technische System der National Library of Australia wird die British Library übernehmen, die in einem Konsortium zusammen mit den National Archives, dem UK Web Archives Consortia, beginnen wird, selektiv britische Websites zu archivieren<sup>5</sup>. Ziel des UK Consortia ist es, in einem ersten Schritt ca. 6 000 Websites auszuwählen und zu archivieren. Zugleich soll getestet werden, inwieweit ergänzend auch eine Archivierung des gesamten britischen Web nach dem Vorbild der schwedischen Nationalbibliothek sinnvoll sein kann, d. h., inwieweit automatisierte Verfahren und intellektuelle Auswahl kombiniert werden können. Gegebenenfalls sollen auch rückwärtige Versionen aus dem Bestand des Internet Archive erworben werden. Daß das Thema Archivierung von Websites auf nationaler Ebene zu einer verstärkten Kooperation von Nationalbibliothek und Nationalarchiv führen kann, zeigt nicht nur das britische Beispiel, sondern anschaulicher noch die Situation in Kanada, wo mittlerweile beide Institutionen vereint sind (zu Library and Archives Canada)<sup>6</sup>.

Den fachlichen Ansatz verfolgt derzeit die Library of Congress in Kooperation mit dem Internet Archive. Ziel der Library of Congress ist es, zunächst zu bestimmten fachlichen Themen, wie den Wahlen oder dem Terroranschlag des 11. September, Archive aufzubauen. Dies erfolgt derzeit technisch zum Teil durch das Internet Archive. Auf mittlere Sicht will die Library of Congress im Rahmen der National Digital Information Infrastructure, für die seit 2000 insgesamt ca. 100 Mio. \$ zur Verfügung stehen, mit weiteren Partnern (ca. 50-75 Partnerinstitutionen sind avisiert) solche thematischen Archive aufbauen, um daran testen zu können, welche Probleme sich für eine Langzeitarchivierung ergeben<sup>7</sup>. Weitere Beispiele für fachliche Archive sind das Digital Archive for Chinese Studies<sup>8</sup> oder das erwähnte 9/11 archive, das von der Georg Mason University, der City University of New York in Kooperation mit der Library of Congress aufgebaut wurde (zur Zeit 57 000 Dateien; 13 GB)<sup>9</sup>. In diesen Kontext gehören auch große digitale Datensammlungen, wie sie zum Beispiel bei geographischen Informationssystemen für die Produktion digitaler Karten anfallen (Bsp.: Geoscience Australia; 500 Terabyte Daten; Produktion von ca. 400 000 Karten [online] pro Jahr)<sup>10</sup>.

Eine weitere im Bereich der Archivierung digitaler Publikationen engagierte Nationalbibliothek ist die Königliche Bibliothek in Den Haag. Sie hat sich bislang auf die Archivierung digitaler Versionen von Verlagszeitschriften konzentriert und bietet dies als Service mittlerweile nicht nur für niederländische Verleger an. Ihr e-depot ist seit 2003 in Betrieb und enthält 3 Mio. Dateien (vor allem PDF-Dateien)<sup>11</sup>. Daß generell die Nationalbibliotheken versuchen, über geänderte Pflichtexemplargesetze (legal deposits) auch digitale Publikationen und Websites zu integrieren, zeigen entsprechende Aktivitäten (z. B. New Zealand; entsprechendes Gesetz seit 2003)<sup>12</sup>. Vorgestellt wurden auch die Projekte der Deutschen Bibliothek, wie Nestor oder Kopal, in denen derzeit an organisatorischen wie technischen Konzepten für die bibliothekarische Langzeitarchivierung in Deutschland gearbeitet wird<sup>13</sup>.

Neben den Strategien und konkreten Projekten der Nationalbibliotheken wurden am dritten Tag vor allem Fragen

der Organisation und der Technik thematisiert. Als eine der international aktivsten Kommissionen kann derzeit das International Internet Preservation Consortium (IIPC) bezeichnet werden, das zur Zeit von der Bibliothèque Nationale in Paris organisiert wird und an dem vor allem Nationalbibliotheken als Mitglieder beteiligt sind<sup>14</sup>. Neben der Entwicklung einschlägiger Standards und der Organisation der Kommunikation der Experten ist das IIPC auch dabei, konkrete technische Produkte zu entwickeln, die für die Langzeitarchivierung eingesetzt werden können.

Abgesehen vom Thema Metadaten wurden ferner einige Projekte vorgestellt, die sich mit softwaretechnischen Fragen befassen, die auftreten, wenn man konkret über die langfristige Archivierung digitaler Dateien nachdenkt. Ein grundlegendes Problem ist dabei, daß für zukünftige Konversionen sichergestellt sein muß, daß die technischen Dokumentationen aller Dateiformate, die in einem Archiv vorhanden sind, überliefert und zugänglich sind. Dazu gibt es derzeit sogenannte File Format Registry-Projekte (unter anderem von der Harvard University oder den National Archives des UK), die versuchen, genau diese Dokumentationen zu sammeln und über standardisierte Schnittstellen zugänglich zu machen<sup>15</sup>. Solche Services sind unabdingbare Voraussetzung nicht nur dafür, daß in Zukunft Konversionsprogramme geschrieben werden können, sondern auch für die Administration von Archivservern. Die Cornell University betreibt z. B. ein Virtual Remote Control Project, bei dem es um die Entwicklung eines „Frühwarnsystems“ für Archivserver geht, das Systemverwaltern meldet, wenn z. B. Dateiformate im Archiv dabei sind zu veralten und konvertiert werden sollten<sup>16</sup>.

<sup>5</sup> UK Web Archiving Consortium <<http://www.webarchive.org.uk/>>.

<sup>6</sup> Library and Archives Canada <<http://www.collectionscanada.ca/index-e.html>>.

<sup>7</sup> Digital Preservation <<http://www.digitalpreservation.gov/>>.

<sup>8</sup> Digital Archive for Chinese Studies <<http://www.sino.uni-heidelberg.de/dachs/leiden/>>.

<sup>9</sup> The September 11 Digital Archive <<http://911digitalarchive.org/>>.

<sup>10</sup> Geoscience Australia <<http://www.ga.gov.au/>>.

<sup>11</sup> KB Den Haag: e-depot and digital preservation <<http://www.kb.nl/dnp/e-depot/e-depot-en.html>>.

<sup>12</sup> Zu den Aktivitäten der neuseeländischen Nationalbibliothek insgesamt vgl. die Pressemitteilung National Library to capture New Zealand's digital heritage <<http://www.natlib.govt.nz/bin/media/pr?item=1085885702>>.

<sup>13</sup> Nestor. Kompetenznetzwerk Langzeitarchivierung <<http://www.langzeitarchivierung.de/>>.

<sup>14</sup> International Internet Preservation Consortium (IIPC) <<http://www.netpreserve.org/about/index.php>>.

<sup>15</sup> Vgl. Steven L. Abrams: Towards a Global Format Registry. In: World Library and Information Congress: 69th IFLA General Conference and Council, August 1-9, 2003, Berlin, Germany <[http://www.ifla.org/IV/ifla69/papers/128e-Abrams\\_Seaman.pdf](http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf)> sowie Global Digital Format Registry (GDFR) <<http://hul.harvard.edu/gdfr/>>; PRONOM. The file format registry <<http://www.nationalarchives.gov.uk/pronom/>>.

<sup>16</sup> Vgl. Nancy Y. McGovern, Anne R. Kenney, Richard Entlich, William R. Kehoe und Ellie Buckley: Virtual Remote Control: Building a Preservation Risk Management Toolbox for Web

Außerhalb der Konferenz gab es als Abschluß noch einen sog. Information Day, bei dem konkrete technische Produkte vorgestellt wurden.

Zunächst wurde natürlich Pandas<sup>17</sup> vorgestellt, das Archivierungssystem der National Library of Australia. Die wesentlichen Workflowfunktionen, welche dieses Systems unterstützt, sind:

- Identifikation, Auswahl und Registrierung einschlägiger Ressourcen (Websites).
- Anfrage um Erlaubnis der Archivierung und Verwaltung dieser Daten.
- Konfiguration von Gatherern zum Einsammeln der Dateien.
- Qualitätskontrolle (Prüfung der Vollständigkeit des Harvestingprozesses etc.).
- Erstellung archivierbarer Datenpakete.
- Organisation des Zugriffs auf die archivierten Versionen von Websites.
- Verwaltung von Nutzerrechten.
- Erfassung administrativer Metadaten.
- Erstellung statistischer Reports etc.

Als weiteres komplettes Archivierungssystem wurde das Nordic Web Archive vorgestellt<sup>18</sup>, das, anders als das australische Modell, keine Eigenentwicklung darstellt, sondern aus Modulen vorhandener Systeme aufgebaut ist (als Crawler wird z. B. Heritrix verwendet [siehe unten]).

An konkreten Tools wurden vorgestellt:

- Heritrix Crawler<sup>19</sup>: eine kooperative Entwicklung des Internet Archives, von IIPC und dem Nordic Web Archive eingesetzt werden soll (und damit als Ersatz für Alexa dienen soll).
- Xinq<sup>20</sup>: ein deep web archiving tool des IIPC, das dazu dient, den Inhalt von Datenbanken auszulesen und in

eine XML-Struktur zu überführen, in der diese Inhalte dann archiviert werden können.

- Xena<sup>21</sup>: eine Entwicklung der National Archives of Australia, das die Formatierung von z. B. Textverarbeitungsdateien (Word) in XML automatisiert durchführt.
- New Zealand Metadata Extractor<sup>22</sup>: ein Tool zur automatischen Erstellung von Metadaten zu Websites.

Insgesamt wurde bei der Konferenz deutlich, daß es bereits eine Reihe von frei nachnutzbaren open source-tools für einige Funktionen des Archivierungsprozesses gibt, die gerade entwickelt werden oder in einer ersten nutzbaren Version im Jahr 2005 vorliegen werden. Zugleich konnte man erkennen, daß in einigen Nationalbibliotheken und -archiven schon seit einigen Jahren konkretes Know-how erworben wurde, wie Langzeitarchivierung konzeptionell und technisch organisiert werden kann. Bezeichnenderweise waren es gerade die Nationalbibliotheken kleinerer Länder (unter der Perspektive der Einwohnerzahl), die hierbei als Vorreiter agierten. Mittlerweile werden aber in Großbritannien wie in den USA konkrete Projekte entwickelt, um auch dort den Aufbau von Archiven von Websites voranzutreiben. Dabei wurde deutlich, daß mittlerweile eine Art Konsens darüber besteht, daß ausgewählte Ressourcen intellektuell selektiert und katalogisiert werden sollten; daß diese Verfahren aber mit automatisierten Verfahren ergänzt und gegebenenfalls kombiniert werden sollten.

#### **Anschrift des Autors:**

Wilfried Enderle  
Niedersächsische Staats- und  
Universitätsbibliothek Göttingen  
D-37070 Göttingen

Resources. In: D-Lib Magazine, April 2004 <<http://www.dlib.org/dlib/april04/mcgovern/04mcgovern.html>>. Oder zu einem vergleichbaren Projekt: The PANIC (Preservation web-services Architecture for Newmedia and Interactive Collections): <<http://metadata.net/panic/>>.

<sup>17</sup> Pandas <<http://pandora.nla.gov.au/pandas.html>>.

<sup>18</sup> Nordic Web Archive <<http://nwa.nb.no/>>.

<sup>19</sup> Heritrix: Internet Archive's open source web crawler <<http://crawler.archive.org/>>.

<sup>20</sup> Xinq – the XML Database Archive Access Tool <<http://sourceforge.net/projects/xinq>>.

<sup>21</sup> Xena <<http://sourceforge.net/projects/xena/>>.

<sup>22</sup> National Library of New Zealand Metadata Extraction Tool Version 1.0 <<http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#extraction>>.