Tanja Friedrich
Staatsbibliothek zu Berlin – Preußischer Kulturbesitz

Webarchivierung an der National Library of Australia

National Library of Australia – Foyer

0 Einleitung

Im Rahmen meiner praktischen Ausbildung zur wissenschaftlichen Bibliothekarin an der *Staatsbibliothek zu Berlin – Preußischer Kulturbesitz* hatte ich Gelegenheit, gefördert von Bibliothek & Information International, vom 24.8.2009 bis zum 11.9.2009 die *National Library of Australia* (NLA) in Canberra zu besuchen und insbesondere deren Abteilung *Web Archiving & Digital Preservation* kennenzulernen.
Ich habe mich für diesen Fachaufenthalt entschieden, um meine im bisherigen Berufsleben erlangten Kenntnisse in der Digitalisierung um den Bereich der Archivierung digitaler Inhalte zu ergänzen. Aufgrund ihrer wegweisenden Aktivitäten auf diesem Gebiet fiel meine Wahl auf die NLA, die bereits seit 1996 das australische Webarchiv PANDORA[1] betreibt und weiterentwickelt.
Während meines dreiwöchigen Aufenthalts an der Nationalbibliothek lernte ich folgende Bereiche näher kennen: *Web Archiving* (Schwerpunkt), *Digital Preservation*, *Collection Delivery and Storage*, *Digitisation & Photography Branch*, *Libraries Australia* und *Collaborative Services*.

1 Die National Library of Australia

Mein Besuch begann mit einem allgemeinen Informationstag in der Personalabteilung und im von mir gewählten Schwerpunktbereich Webarchivierung. Ich wurde an diesem Tag mit der Verfaßtheit und Struktur der NLA vertraut gemacht. Die NLA ist eine nachgeordnete Behörde des Ministeriums für Umwelt, Wasser, Kulturelles Erbe und Kunst (*Department of the Environment, Water, Heritage and the Arts*)[2] und als solche eine Körperschaft Öffentlichen Rechts (*body corporate*)[3]. Sie hat mehr als 470 Mitarbeiterinnen und Mitarbeiter und verfügt über ein Haushaltseinkommen (Etat + Einkünfte) in Höhe von 62 Millionen Australische Dollar, wovon mehr als die Hälfte für Personalkosten aufgewendet werden.[4] Der Bestand beläuft sich auf rund 9 Millionen Medieneinheiten, darin enthalten auch Photographien (etwa 800.000), Digitalisate (etwa 146.000) und Tonaufnahmen (etwa 17.000).[5] Die NLA ist damit die größte Bibliothek Australiens.
In ihrer Funktion als Nationalbibliothek sammelt und bewahrt die NLA Materialien mit Bezug zu Australien und zur australischen Bevölkerung. Sie hat das Pflichtexemplarrecht für alle australischen Printpublikationen inne.[6] Daneben verfügt die Bibliothek über zahlreiche Überseepublikationen (darunter auch etwa 70.000 Rara[7]) und einen beachtlichen Bestand an asiatisch-pazifischen Publikationen. In der NLA befinden sich außerdem große Sondersammlungen in den Bereichen Zeitungen, Handschriften, Karten, Musik und Tanz, Mündliche Überlieferung und Folklore. Eine besondere und bedeutende Sondersammlung stellt das an der NLA geführte Bildarchiv (*NLA Pictures Collection*) dar.
Zum Verantwortungsbereich der NLA gehört auch das Angebot verschiedener Serviceleistungen für australische Bibliotheken, zum Beispiel der Betrieb des nationalen Verbundkatalogs *Libraries Australia*. Diese

[1] Das Akronym PANDORA steht für die Zielsetzung des australischen Webarchivs: *Preserving and Accessing Networked Documentary Resources of Australia*. Der vollständige Name des Webarchivs lautet *PANDORA Australia's Web Archive*.
[2] *Service Charter*. URL: http://www.nla.gov.au/charter/.
[3] *National Library Act 1960 Sect 5*. URL: http://scaletext.law.gov.au/html/pasteact/1/761/0/PA000090.htm.
[4] *General Facts about the National Library of Australia*. URL: http://www.nla.gov.au/library/factsheet.html.
[5] Ebda. Der Anteil an Monographien beträgt etwa 3 Millionen, an Handschriften etwa 2 Millionen.
[6] *Copyright Act 1968 Sect. 201*. URL: http://www.austlii.edu.au/au/legis/cth/consol_act/ca1968133/s201.html.
[7] *What we collect. Rare Books*. URL: http://www.nla.gov.au/collect/rarecoll.html.

Dienstleistungen für Bibliotheken werden in der Abteilung *Resource Sharing & Innovation* erbracht, einer der drei bibliothekarischen Hauptabteilungen. Insgesamt hat die NLA sieben Abteilungen:

1. *Collections Management*
2. *Australian Collections & Reader Services*
3. *Resource Sharing & Innovation*

(= drei bibliothekarische Hauptabteilungen)

4. *Information Technology*

5. *Corporate Services*
6. *Executive Support & Public Programs*
7. *Communications, Marketing & Community Programs*

(= drei Verwaltungsabteilungen)

Die mitarbeiterstärkste Abteilung ist die bibliothekarische Kernabteilung *Collections Management*. Insbesondere für die Einarbeitung der Zeitschriften und Serien wird viel bibliothekarisches Personal beschäftigt. Ein weiterer großer Bereich dieser Abteilung ist die Monographienbearbeitung, die durch das Pflichtexemplarrecht ebenfalls mit hohem Ressourcenaufwand betrieben werden muß. Weitere Bereiche des *Collections Managements* sind *Bibliographic Standards & Strategy*, *Collections Preservation*, *Digital Collections Management* und auch der Bereich *Web Archiving & Digital Preservation*, in dem ich in den drei Wochen nach einleitendem Training selbständig mitarbeiten durfte. Ein weiterer, siebter Bereich der Abteilung ist *Asian Collections*, dem an der NLA ein besonderer Stellenwert zukommt. Aufgrund besonderer Schriftzeichen erfolgt die Katalogisierung dieser Sammlung in einem eigenen Team des *Collections Managements*. Die anderen Sondersammlungen der Bibliothek werden dagegen von der allgemeinen Katalogisierung (Bereiche *Serials* und *Monographs*) mitbetreut. Der Aufbau der Sondersammlungen (Handschriften, Karten, Musik und Tanz, Mündliche Überlieferung und Folklore, Bildarchiv) und die Vermittlung des gesamten Bestands wird in der zweiten bibliothekarischen Hauptabteilung *Australian Collections & Reader Services* geleistet. In dieser Abteilung sind außerdem die Magazinverwaltung und die Bereitstellung angesiedelt. Die Betreuung der Leserinnen und Leser erfolgt hauptsächlich im allgemeinen Lesesaal, in den Lesesälen der Sondersammlungen sowie im Forschungslesesaal *Petherick Reading Room*, der für Nutzerinnen und Nutzer mit besonderem Forschungsinteresse reserviert ist. Alle Leserinnen und Leser können Beratung zu ihren Forschungsfragen und zu den Beständen bei den *Reference Librarians*[8] erhalten. Begründet in der Geschichte Australiens besteht in der Nutzerschaft der Bibliothek ein besonderes Interesse an genealogischer Forschung. Die *Reference Librarians* können zwar keine familiengeschichtlichen Forschungsaufträge ausführen, geben aber täglich vielen Interessierten Auskunft zu genealogischen Methoden und Hilfsmitteln. Es stehen zahlreiche spezifische Informationsmittel für die Familienforschung zur Verfügung, die in einem eigenen Bereich der NLA-Internetseiten vorgestellt und erläutert werden.[9]

[8] Australische *Reference Librarians* oder Auskunftsbibliothekarinnen und -bibliothekare sind, ähnlich wie in den USA, Expertinnen und Experten für die mündliche und schriftliche Auskunft für Leserinnen und Leser in Bibliotheken.
[9] *For family historians*. URL: http://www.nla.gov.au/infoserv/family.html.

In den kommenden drei Wochen war ich wunschgemäß im Bereich *Web Archiving* der Abteilung *Collections Management* untergebracht. In halb- bis ganztägigen Informationsbesuchen lernte ich zusätzlich zwei weitere Sektionen dieser Abteilung kennen: *Digital Preservation* und den Bereich *Digitisation & Photography Branch* (DAP). Auch in den anderen beiden bibliothekarischen Hauptabteilungen konnte ich Informationsgespräche mit Verantwortlichen führen: In der Abteilung *Australian Collections & Reader Services* zum Bereich *Collection Delivery and Storage* und in der Abteilung *Resource Sharing & Innovation* zu *Libraries Australia* und *Collaborative Services*.

2 Webarchivierung an der NLA

2.1 Allgemeines

Die Webarchivierung und die digitale Langzeitarchivierung sind an der NLA zusammengefaßt im Bereich *Web Archiving and Digital Preservation* in der Abteilung *Collections Management*. Der Teilbereich *Web Archiving* wird von Paul Koerbin geleitet, in dessen Team ich mitarbeiten durfte. Die Bibliothek stellte mir einen Computerarbeitsplatz zur Verfügung, sodaß ich unabhängig und selbständig arbeiten konnte. Mir wurde sehr zügig und unkompliziert Zugang zu den wesentlichen Arbeitsinstrumenten ermöglicht. So hatte ich von Beginn an ein Emailkonto, Zugang zum Schriftgutverwaltungssystem[10], Zugriff auf das Intranet, das interne Wiki, die für meine Arbeit relevanten Netzlaufwerke sowie ein Benutzerkonto für die Webarchivierungssoftware PANDAS[11].

Im Bereich *Web Archiving* sind derzeit vier Personen beschäftigt, Paul Koerbin als Leiter und drei sogenannte Webkuratoren (*Web Curators*), die die Webarchivierung durchführen. Angesichts der Aufgabe, Internetseiten zu archivieren, deren Inhalt sich mit Australien beschäftigt, die von Australiern erstellt wurden oder die ansonsten von gesellschaftlicher, politischer, kultureller, religiöser, wissenschaftlicher oder wirtschaftlicher Bedeutung für Australien sind,[12] erscheint die Personalausstattung nicht gerade groß. Vor diesem Hintergrund ist auch die Zahl von etwa 22.300[13] seit 1996 im Archiv veröffentlichten Titeln zu sehen, die gering erscheinen mag. Allerdings hängt der Umfang des Archivs auch mit der gewählten selektiven Archivierungsmethode zusammen, die im weiteren Verlauf dieses Berichts näher erläutert wird. Die archivierten Internetseiten werden der Öffentlichkeit in dem international bekannten und renommierten australischen Webarchiv PANDORA zur Verfügung gestellt. Dort können die Titel mit einer Volltextsuche recherchiert werden.

Seit 1998 arbeitet die NLA beim Aufbau von PANDORA mit weiteren Institutionen zusammen, die jeweils für die Archivierung von Onlinepublikationen, die ihren regionalen oder thematischen Zuständigkeitsbereich betreffen, verantwortlich sind. Mittlerweile sind neun Partnerinstitutionen am Aufbau von PANDORA beteiligt: *The Northern Territory Library*, *The State Library of Victoria*, *The State Library of New South Wales*, *The State Library of Queensland*, *The State Library of Western Australia*, *The State Library of South Australia*, *The National Film and Sound Archive*, *The Australian War Memorial* und *The Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS)*.

---

[10] In der NLA und anderen australischen Behörden wird für die Verwaltung von Vorgängen ein elektronisches Schriftgutverwaltungssystem oder *Records Management* verwendet. Dieses dient zum Beispiel zur institutionsübergreifenden Archivierung von Emailkorrespondenz.

[11] PANDAS = *PANDORA Digital Archiving System*. Weiterführende Informationen zu PANDAS unter http://pandora.nla.gov.au/pandas.html. Zu PANDORA s. Anm. 1.

[12] Vgl. *The purpose of the PANDORA archive*. URL: http://pandora.nla.gov.au/overview.html#purposearchive.

[13] Die aktuelle Zahl veröffentlichter Archivpublikationen ist abrufbar unter http://pandora.nla.gov.au/alpha/ALL.

In diesen Institutionen, die im PANDORA-Kontext meistens als Agenturen (*agencies*) bezeichnet werden, arbeitet in der Regel eine Person an der Auswahl und Archivierung der Publikationen. Die NLA unterstützt diese Mitarbeiterinnen und Mitarbeiter beim Umgang mit der Workflowsoftware PANDAS. Die Nationalbibliothek ist außerdem für das Hosting des gesamten Archivs und die Weiterentwicklung der Archivierungssoftware und -prozesse verantwortlich.
Meine Mitarbeit im Bereich *Web archiving* nahm den größten Teil meiner Praktikumszeit ein. Als temporäres Teammitglied war ich in den täglichen Arbeitsablauf eingebunden und lernte die Tätigkeit der Webarchivierung praktisch kennen.

2.2 Geschichte der Webarchivierung an der NLA

In ihrer Funktion als Nationalbibliothek ist die NLA in der Verantwortung, Materialien unterschiedlichster Formate zu sammeln, die die Geschichte und Kultur Australiens repräsentieren. Die Sammlungsaktivität der NLA erstreckt sich konsequenterweise nicht nur auf Bücher, Zeitschriften und Zeitungen, sondern zum Beispiel auch auf Bilder, Karten oder Tondokumente. Bereits in den 1990er Jahren erkannte die Bibliothek, daß mit dem immer schneller an Bedeutung gewinnenden Internet wichtige australische Inhalte in einem weiteren neuen Format Verbreitung fanden: als Webseiten. Schon in den frühen Jahren des *World Wide Web* identifizierte die NLA damit als eine der ersten Bibliotheken die Problematik der Einzigartigkeit von Informationen im neuen Medium bei gleichzeitiger Nicht-Persistenz. Ihre Verantwortung, Materialien australischen Inhalts zu sammeln, um historisch und kulturell bedeutende Information zugänglich zu machen und zu bewahren, bezieht sie seither auch auf Online-Inhalte, auch wenn diese vom Pflichtexemplarrecht bislang nicht erfaßt sind.
Die NLA beschloß angesichts der Menge und technischen Komplexität der potentiell relevanten Online-Informationen, eine Sammlung von Webpublikationen nicht im Alleingang, sondern in Kooperation mit Partnerinstitutionen anzugehen, allen voran mit den australischen *State Libraries*. Die Grundlagen für dieses Vorhaben wurden zunächst mit der Entwicklung der notwendigen Technik in der Nationalbibliothek geschaffen. 1996 archivierte die NLA damit die ersten beiden Titel für das geplante australische Webarchiv PANDORA. Zwei Jahre später waren Infrastruktur und Vorgehensweise stabil und praktikabel genug, um die *State Libraries* zur Mitwirkung einzubeziehen.

2.3 Technik der Webarchivierung

Archivieren von Webseiten bedeutet im wesentlichen, Kopien von im Internet veröffentlichten Seiten zu erstellen und sie in genau der Version, in der sie zu diesem Zeitpunkt vorliegen, auf einem Archivserver abzuspeichern. In dieser Version sollen sie möglichst originalgetreu in der Archivumgebung abrufbar sein und bleiben, woraus sich zwei zentrale Herausforderungen für die Webarchivierung ergeben:

a. Die Funktionalitäten und vor allem Inhalte der Originalseite sollten vollständig abgebildet werden, und

b. die Daten sollten in einer Form archiviert werden, die den Zugriff auf lange Zeit sichert.

Diese beiden zentralen Herausforderungen stehen in einem gewissen Spannungsverhältnis zueinander. Einerseits müssen für die originalgetreue Abbildung der Webseiten Dateien in den unterschiedlichsten Formaten übernommen werden. Durch neue Entwicklungen in der Webprogrammierung nimmt die Komplexität der Funktionalität der Seiten außerdem immer weiter zu. Andererseits ist es aus der Perspektive der Langzeitarchivierung erstrebenswert, mit bewährten Formaten zu arbeiten und gewissen Standards zu folgen.

Um diesem Dilemma möglichst effektiv begegnen zu können, arbeiten die beiden Bereiche *Web Archiving* und *Digital Preservation* an der NLA eng zusammen.

2.3.1 Domain harvesting

Prinzipiell gibt es zwei unterschiedliche Herangehensweisen bei der Archivierung von Webseiten. Die eine, unter anderem vom *Internet Archive*[14] verwendete Methode ist das sogenannte *Domain harvesting*[15]. Bei diesem Verfahren wird ein *Harvester*[16] darauf programmiert, alle in einem bestimmten Zeitraum (zum Beispiel von zwei Wochen) erreichbaren Webseiten zu kopieren, die auf einer bestimmten Internetdomain, zum Beispiel .com, online sind. Der *Harvester* wird dabei üblicherweise so eingestellt, daß er entweder für eine bestimmte Zeit auf einer Webpräsenz verweilt und dann, unabhängig davon, ob er bereits alle dort vorhandenen Dateien gespeichert hat, zur nächsten weiterzieht. Oder er speichert die Inhalte immer nur bis zu einer bestimmten Tiefe, zum Beispiel bis zur dritten Hierarchieebene der Internetseiten der gewünschten Domain. Dieses *Domain harvesting* hat den Vorteil, daß innerhalb kurzer Zeit sehr viele Inhalte gespeichert werden können, natürlich unter der Prämisse eines immensen Speicherplatzbedarfs. Ein Nachteil ist, daß bei der Menge der anfallenden Daten keine echte Qualitätskontrolle durchgeführt werden kann. Entsprechend findet man in Webarchiven, die auf *Domain harvests* beruhen, häufig Seiten, die nur zum Teil oder gar nicht funktionieren.

2.3.2 Gezieltes Harvesting

Die Alternative zum *Domain harvesting* ist ein gezieltes *Harvesting* von ausgewählten Internetseiten. Diese Vorgehensweise ist ungleich aufwendiger, ermöglicht aber eine Qualitätskontrolle und ein differenziertes Datenmanagement. Im Vergleich zur riesigen Anzahl an Webseiten, die in einem *Domain harvest* gewonnen werden, kann beim gezielten *Harvesting* sehr viel weniger archiviert werden. Der geringeren Ausbeute wird aber zumindest teilweise dadurch begegnet, daß eine tatsächliche Selektion der Inhalte stattfindet und so von vornherein keine Energie in die Archivierung von Seiten investiert wird, deren Speicherung nicht wünschenswert ist. Natürlich ist der Selektionseffekt nur in solchen Vorhaben nützlich, in denen ein bestimmter inhaltlicher Fokus definiert ist. Das *Internet Archive* geht beispielsweise nicht inhaltlich fokussiert vor, sondern möchte das Web möglichst in seiner Breite archivieren.

2.4 Selektives Archivieren und Domain harvesting an der NLA

Die NLA speichert im australischen Webarchiv PANDORA Internetseiten mit der Methode des selektiven, gezielten *Harvestings*. Hierfür gibt es mehrere Gründe. Erstens verfolgt PANDORA den Anspruch, eine möglichst gute Funktionalität des Archivmaterials zu liefern. Zweitens wird trotz Volltextdurchsuchbarkeit Wert auf eine formale und inhaltliche Erschließung der archivierten Seiten gelegt, um zum Beispiel Sammlungen zu aktuellen (Forschungs-)Themen bilden zu können. Drittens sollen die begrenzten Ressourcen möglichst für die Archivierung solcher Seiten eingesetzt werden, von denen angenommen wird, daß sie für die Dokumentation und Erforschung der australischen Kultur heute und in Zukunft von Bedeutung sind. Im Wissen darum, daß dieser Anspruch nie vollständig erfüllt werden kann – wer weiß schon, was künftig für die Forschung von Interesse ist – hat die NLA 2005 damit begonnen, sich auch in der Technik des *Domain harvestings* eine gewisse

[14] Das *Internet Archive* ist eine US-amerikanische Non-Profit-Organisation, die digitale Inhalte, unter anderem auch Webseiten, mit dem Ziel der Langzeitarchivierung speichert und im Internet zur Verfügung stellt. URL: http://www.archive.org/index.php.
[15] Mit dem englischen Begriff *harvesting* (dt. ‚ernten') ist im hier behandelten Zusammenhang das softwaregesteuerte Aufrufen und Abspeichern (oder Herunterladen) von Internetseiten gemeint.
[16] Eine andere Bezeichnung für eine zum *Domain harvesting* verwendete Software ist *Web Crawler*.

Expertise anzueignen. Seither wird neben dem gezielten *Harvesting* vorerst jährlich in Zusammenarbeit mit dem *Internet Archive* als Dienstleister ein *Harvesting* der gesamten Domain .au durchgeführt.[17] Die dabei gespeicherten Daten werden aus den USA auf sogenannten *PetaBoxes*[18] geliefert und in der NLA in dieser Form archiviert. Die Daten aus den *Domain harvests* werden allerdings nicht zusammen mit den anderen Daten in PANDORA der Öffentlichkeit zur Verfügung gestellt. Abgesehen von technischen Gründen liegt das vor allem daran, daß die NLA kein Pflichtexemplarrecht für Webinhalte hat. Wollte sie die Daten in PANDORA veröffentlichen, müßte sie zunächst bei allen Herausgebern die Erlaubnis für die Veröffentlichung nach *Copyright Act 1968*[19] einholen.

2.4.1 Selektionskriterien der NLA und ihrer Partneragenturen

Eine selektive Webarchivierung, die dem Anspruch genügen will, zur Dokumentation der Geschichte und Kultur Australiens beizutragen, erfordert ein gezieltes Vorgehen anhand kritisch formulierter Selektionskriterien. Der Aufwand, der mit dem selektiven *Harvesting* verbunden ist, ist nur dann gerechtfertigt, wenn die archivierten Inhalte tatsächlich eine gewisse Bedeutung haben, wenn möglich auch für künftige Forschungsinteressen.
Die NLA und die anderen neun Partneragenturen haben jeweils eigene Selektionskriterien erarbeitet, nach denen sie bei der Auswahl der zu archivierenden Artikel vorgehen.
Grundlegend orientiert sich die Auswahl am allgemeinen Erwerbungsprofil der jeweiligen Einrichtung, weshalb es eine regional verteilte und eine thematische Zuständigkeit der einzelnen Agenturen gibt.[20] Im Fokus der NLA stehen Webseiten mit überregional relevantem Inhalt und mit Bezug zum Australian Capital Territory (ACT).
Die *State Libraries* sind für die Archivierung der Seiten verantwortlich, die relevant für ihren Staat bzw. für ihr Territorium sind. *ScreenSound Australia* wählt Internetseiten mit Bezug zu australischer Musik und australischem Film aus, während sich das *Australian War Memorial* für Inhalte zur australischen Militärgeschichte in der Verantwortung sieht. Schließlich behandelt das *Australian Institute of Aboriginal and Torres Strait Islander Studies* Publikationen zu Themen der indigenen Völker Australiens.
Über regionale und thematische Kriterien hinaus hat jede Agentur weitere Relevanzkriterien definiert, die bei der Auswahl der zu archivierenden Titel herangezogen werden.
Die Relevanzkriterien der NLA beziehen sich zum Beispiel auf die Herausgeberschaft. So werden mit Priorität unter anderem Regierungsseiten oder Publikationen von Bildungseinrichtungen Australiens archiviert. Seiten mit kommerziell werbendem Inhalt sind dagegen beispielsweise ausgeschlossen. Nicht archiviert werden im Übrigen auch solche Titel, die auch im Printformat zur Verfügung stehen, etwa auch im Internet veröffentlichte Konferenzschriften o.ä. Eine Besonderheit im Selektionsprofil der NLA ist außerdem die Auswahl von Webseiten zu für eine Laufzeit von drei Jahren bestimmten Schwerpunktthemen und zu aktuellen, gesellschaftlich bedeutenden Ereignissen[21].

---

[17] Informationen zur Vorgehensweise und quantitative Analysen zu den *Domain harvests* liegen vor in: Koerbin, Paul (2008): *The Australian web domain harvests: a preliminary quantitative analysis of the archive data*. URL: http://pandora.nla.gov.au/documents/auscrawls.pdf.
[18] *PetaBoxes* sind Speicherplatten, die für besonders große Archivierungsbedürfnisse verwendet werden können. Webseite des Anbieters: http://www.capricorn-tech.com/.
[19] Australisches Copyright-Gesetz, verfügbar unter http://www.austlii.edu.au/au/legis/cth/consol_act/ca1968133/.
[20] Die Selektionskriterien aller Institutionen sind abrufbar unter http://pandora.nla.gov.au/selectionguidelinesallpartners.html.
[21] So gibt es zum Beispiel entsprechend erschlossene Sammlungen zu den Themen *Sydney Olympics* (http://pandora.nla.gov.au/col/4006) oder *Bali bombing* (http://pandora.nla.gov.au/col/8200).

2.4.2 Webarchivierung mit PANDAS

Zur technischen Durchführung der selektiven Webarchivierung hat die NLA mangels geeigneter Alternativen in Eigenleistung ein Softwaresystem entwickelt: das *PANDORA Digital Archiving System* (PANDAS)[22]. Wichtig war dabei, eine Lösung zu erhalten, mit der alle Einrichtungen kooperativ arbeiten können, weshalb PANDAS unbedingt ein webbasiertes Programm sein sollte, das ein komfortables Datenmanagement ermöglicht. Die aktuelle Version 3 wurde 2007 in Betrieb genommen.

In PANDAS kann für alle zu archivierenden Titel ein Eintrag erstellt werden, und jedem Titel wird ein verantwortlicher Bearbeiter (*Web Curator*) der zuständigen Institution zugeordnet. Alle Titel sind in PANDAS im jeweiligen Arbeitsstadium sichtbar. Das Arbeitsstadium richtet sich nach den Arbeitsschritten die im Archivierungsprozeß bereits durchgeführt wurden. Für jedes Arbeitsstadium gibt es in PANDAS einen eigenen Arbeitsbereich (*Work tray*), in dem die Titel angezeigt werden, die für den nächsten Arbeitsschritt bereit sind. Jeder *Web Curator* sieht nur diejenigen Titel, für die er selbst zuständig ist. Damit ist das System sehr übersichtlich, und die einzelnen Arbeitsvorgänge können gut geplant werden.

Die PANDAS-Arbeitsschritte sind im einzelnen: Vorschlag, Auswahl, Erlaubnis des Herausgebers, Herunterladen, Qualitätskontrolle, Archivieren, Katalogisieren und Veröffentlichen. Diese Arbeitsschritte bauen aufeinander auf. Bei der erfolgreichen Bearbeitung eines Schritts wird der Titel durch das System in den nächsten Arbeitsbereich (*Work tray*) verschoben. Kann ein Schritt nicht erfolgreich bearbeitet werden, besteht die Möglichkeit, den Titel als nicht archivierbar zu kennzeichnen, wodurch er zwar nicht aus dem System gelöscht wird, aber nicht mehr im Arbeitsbereich des *Web Curators* auftaucht.

2.5 Zukunft der Webarchivierung an der NLA

Die NLA gehört zu den Bibliotheken, die schon sehr früh erkannt haben, daß im Internet nicht nur unikale Inhalte von großer Relevanz für die Forschung in immer größerer Zahl erscheinen, sondern daß diese Inhalte auch häufig verschwinden oder nicht mehr abrufbar sind und damit für die Forschung verloren gehen. Ebenfalls hat die NLA – sich ihrer Verantwortung für diese Inhalte bewußt – von Anfang an gesehen, daß sie diese Herausforderung nur zusammen mit Kooperationspartnern würde annehmen können.

Seit dem Start von *PANDORA Australia's Web Archive* sind inzwischen 13 Jahre vergangen und die Herausforderungen in der Webarchivierung sind seither in allen Dimensionen angestiegen. Das Internet wächst permanent und mit ihm die potentiell forschungsrelevanten, unikalen Publikationen darin. Gleichzeitig wird das Internet dynamischer und die Technik ändert sich ständig, was das Herunterladen und Sichern der Informationen zu einer immer komplexeren Aufgabe macht.

Wie kann und will die NLA diesen wachsenden Herausforderungen begegnen? Eine Antwort könnte sein, sich von der selektiven Archivierung zu verabschieden und alle Ressourcen in ein regelmäßiges, umfassendes *Domain harvesting* zu investieren. Wenn auch auf Kosten der Qualität ließen sich so wesentlich mehr Inhalte sichern und der Öffentlichkeit zur Verfügung stellen. Voraussetzung hierfür wäre allerdings eine Ausweitung des australischen Pflichtexemplarrechts auf Webpublikationen. Nur damit könnten Webseiten ohne explizites Einverständnis der Herausgeber archiviert und veröffentlicht werden. Allerdings gibt es außer den Copyrightproblemen weitere Rechtsbereiche, derentwegen eine Selektion immer notwendig bliebe, etwa Fragen der Verletzung der Privatsphäre, der Beleidigung oder generell krimineller Inhalte. Schließlich würde die NLA diese Inhalte auf ihren Servern vorhalten und zur Verfügung stellen. Zu Bedenken bliebe außerdem, daß *Domain harvests* sehr kostenintensiv sind, nicht zuletzt wegen der hohen Speicherplatzkapazitäten, die sie erfordern.

Außerdem ist ein rein auf *Domain harvests* beruhendes Webarchiv kaum auf Qualität überprüfbar. Und liegen archivierte australische Webseiten dieser Qualität nicht schon in ausreichender Weise im *Internet Archive* und im *Google*-Index vor? Die NLA verfolgt beim Aufbau all ihrer Sammlungen dagegen das Ziel, die Publikationen für Nutzerinnen und Nutzer nicht nur irgendwie verfügbar zu machen, sondern dies erstens effizient zu tun (Kosten-Nutzen-Relation) und zweitens einen Mehrwert durch ihre Informationsinfrastruktur zu schaffen. Ein Verzicht auf Qualitätsstandards zugunsten einer unbeherrschbaren Masse an Informationseinheiten erscheint vor diesem Hintergrund als ein nicht gangbarer Weg.[23]

Was wäre aber davon zu halten, wenn es ein entsprechendes Pflichtexemplarrecht gäbe und die NLA keine *Harvests* durchführte, sondern die Herausgeber ihre Publikationen aus eigener Initiative ablieferten? Eine weit größere Menge als mit der bisherigen Methode könnte so erfaßt werden, und das aufwendige, häufig erfolglose Ersuchen um Archivierungserlaubnis bliebe erspart. Jedoch käme damit der ungleich größere Aufwand für das *Web Archiving Team* hinzu, die dann viel zahlreicheren neuen Publikationen mit PANDAS zu bearbeiten. Und wäre dieser Aufwand tatsächlich für alle abgelieferten Publikationen gerechtfertigt? Was geschähe außerdem mit den zu erwartenden Mengen an nicht abgelieferten Daten? Wie könnten diese Ausfälle kontrolliert werden? Edgar Crook, Mitarbeiter des *Web Archiving Teams*, hat in einem Vortrag[24] anläßlich der *ALIA*[25] *Biennial Conference* in Alice Springs im September 2008 aufgrund der hier angesprochenen und noch weiterer Gedanken eine andere Zukunft für die Webarchivierung an der australischen Nationalbibliothek aufgezeigt. Er kommt zu dem Ergebnis, daß zwar die derzeitige Vorgehensweise an ihre Grenzen gestoßen sei, und das PANDAS-System aufgrund mangelnder Entwicklungskapazitäten an der NLA wohl nicht mehr weiter an Veränderungen in der Internetwelt angepaßt werden könne. Jedoch zieht er den Schluß, daß gerade aufgrund technologischer Veränderungen und wachsender Inhalte im Netz die Identifizierung der relevanten Inhalte und die Bearbeitung derselben immer wieder notwendig sein werde. Wie das Internet selbst müsse auch die Archivierungstechnik immer variabel und entwickelbar bleiben. Eine Perspektive jenseits PANDAS sieht Crook daher im Engagement der NLA im *International Internet Preservation Consortium* (IIPC)[26]. Dieses Konsortium von in der Webarchivierung aktiven Bibliotheken und anderen Institutionen arbeitet gemeinsam an der Entwicklung neuer Lösungen, die eine effiziente Webarchivierung in Zukunft ermöglichen sollen. Die NLA bringt sich aktiv in diese Bemühungen ein und kann einerseits ihre Erfahrung aus 13 Jahren einbringen und andererseits von den in Form von Softwarelösungen entstehenden Ergebnissen der IIPC-Aktivitäten für die eigene zukünftige Webarchivierungspraxis profitieren. Mit diesem Engagement verfolgt die NLA weiter den Weg, den sie von Anfang an in der Webarchivierung in Kooperation mit den *State Libraries* und anderen gegangen ist und von dessen Unausweichlichkeit man hier überzeugt ist: Die partnerschaftliche Suche nach Lösungen und die kooperative Zusammenarbeit.

---

[22] Weiterführende Informationen zu PANDAS unter http://pandora.nla.gov.au/pandas.html.

[23] Zumal der große Umfang auch eher auf einer breiten (horizontalen), statt auf einer detaillierten (vertikalen) Fülle beruht: Es werden bei den *Domain harvests* zwar viele Webpublikationen gesammelt, diese aber jeweils nur bis zu einem bestimmten Umfang, zum Beispiel die Seiten einer Webpräsenz nur bis zu einer bestimmten Hierarchieebene (s. Abschnitt 2.3.1).

[24] Crook, Edgar (2008): *Web Archiving in a Web 2.0 World. Paper presented at the ALIA Biennial Conference, Alice Springs, on 2 September 2008*. URL: http://pandora.nla.gov.au/pan/13910/20080930-1156/conferences.alia.org.au/alia2008/pdfs/124.TT.pdf.

[25] ALIA = *Australian Library and Information Association.*

[26] Internetseiten des Konsortiums: http://www.netpreserve.org/about/index.php.