World Library and Information Congress:
70th IFLA General Conference and Council

22-27 August 2004
Buenos Aires, Argentina

Programme: http://www.ifla.org/IV/ifla70/prog04.htm

Code Number: 009-G
Meeting: 89. Cataloguing
Simultaneous Interpretation: -

Das Paradigma-Projekt und seine Suche nach Metadatenlösungen und Benutzerdienstleistungen

Carol Van Nuys, Ketil Albertsen, Linda Pedersen et Asborg Stenstad

Das Paradigma-Projekt, Die Nationalbibliothek von Norwegen
carol.vannuys@nb.no, ketil.albertsen@nb.no, linda.pedersen@nb.no, asborg.stenstad@nb.no

Translation by Ulrike Junger

Kurzfassung:

Das Paradigma-Projekt der Nationalbibliothek von Norwegen soll eine zufriedenstellende Pflichtexemplarabgabe aller Arten von digitalen Dokumenten - auch den Millionen von Dokumenten, die im Internet in Norwegen vorhanden sind, sicherstellen. Norwegen wird hoffentlich in der Lage sein, sein digitales Kulturerbe für die Zukunft zu erhalten, indem es Forschern den Zugang zum Internet-Archiv z.B. mittels Metadaten und Volltextsuche gewährt. Dieses Papier gibt eine kurze Beschreibung des Projektes selbst, bevor die Probleme diskutiert werden, denen man bei der Suche nach Metadaten-Standards für das Auffinden von Dokumenten, die Langzeitarchivierung usw. begegnet. Vorgestellt werden die Anwendung der FRBR-Entitäten work, expression, manifestation und item im Design des Archivs ebenso wie Ideen für zukünftige Dienstleistungen: einen Verifizierungs- und Authentifizierungsservice sowie einen Zuteilungsservice für Identifier – beide verfügbar über das Internet.

1 Einführung

1.1 Web-Archivierung in anderen Ländern

Digitale Dokumente verschwinden täglich, und eine Studie[1] zeigt, dass nur 20 % der im Netz gefundenen Dokumente – unverändert – nach einem Jahr dort verblieben sind. Folglich sind

[1] Mannerheim, Johan. The WWW and our digital heritage [online]. – URL: http:://ifla.org/IV/ifla66/papers/158-157e.html

die Möglichkeiten für neue Generationen von Lesern, die heutigen digitalen Dokumente in der Zukunft betrachten zu können, gering. Der Erhalt unseres digitalen Kulturerbes ist eine zunehmend bedeutende und schwierige Aufgabe, und die Nationalbibliothek von Norwegen[2] ist nur eine von vielen Institutionen, die systematisch an der Lösung der damit in Verbindung stehenden juristischen, technischen und bibliografischen Probleme arbeiten.

Ein Teil dieser Aufgabe ist die Sammlung und Archivierung von Dokumenten aus nationalen Internet-Domänen. Verschiedene Länder haben unterschiedliche Sammelstrategien entwickelt: Dänemark [1] und Australien [2] haben sich für das selektive Verfahren entschieden, während Schweden [3], Island und Finnland ihr gesamtes nationales Web-Netz eingesammelt haben. Norwegen ist eine von wenigen Bibliotheken in Europa, die mit ihrem nationalen Internet-System digitale Dokumente erfasst und archiviert, basierend auf dem bestehenden Gesetz zur Pflichtexemplarabgabe[3].

1.2 Das Gesetz zur Pflichtexemplarabgabe

Zweck des Gesetzes zur Pflichtexemplarabgabe ist es [5]:

„[...] sicherzustellen, dass Dokumente, die allgemein verfügbare Informationen enthalten, in nationalen Sammlungen hinterlegt werden, so dass diese Daten des kulturellen und gesellschaftlichen Lebens in Norwegen erhalten bleiben und als Quellenmaterial für Forschungs- und Dokumentationszwecke zur Verfügung gestellt werden können."(§1) Als äußerst fortschrittlich angesehen, als es im Jahre 1989 erlassen wurde, umfasst das Gesetz zur Pflichtexemplarabgabe alle *allgemein zugänglichen* norwegischen Dokumente, gleichgültig auf welchem Medium sie gespeichert sind: z.B. auf Papier, Mikroformen, Fotos, Medienkombinationen, Tonträgern, Filmen, Videos, digitalen Dokumenten und Fernsehprogrammen. Dokumente, die im Ausland für norwegische Verleger veröffentlicht - und solche, die speziell für die norwegische Öffentlichkeit überarbeitet wurden, sind ebenfalls abgedeckt.

Selbstverständlich gab es das World Wide Web im Jahr 1989 noch nicht im Internet. Digitale Dokumente – meistens in der Form von Datenbanken – waren selten im Vergleich zu den Millionen von Internet-Veröffentlichungen in der heutigen Zeit, aber sie waren technisch noch schwierig zu behandeln. Heute kann der Langzeitaufbewahrungsspeicher der Nationalbibliothek 100 TByte Daten speichern; das entspricht wirklich einer sehr großen Zahl von digitalen Dokumenten.

2 Das Paradigma-Projekt

Die Nationalbibliothek von Norwegen startete das Paradigma-Projekt[4] im August 2001. Ziel des Projekts ist es, eine zufriedenstellende Pflichtexemplarabgabe aller norwegischen digitalen Dokumente zu gewährleisten, dazu gehört die Entwicklung der Technologie, der

---

(Überprüft am 15. April 2004)

[2] Für weitere Informationen über die Nationalbibliothek von Norwegen siehe unter URL: http://www.kb.nl/gabriel/libraries/pages generated/no en.html (Überprüft am 15. April 2004)

[3] Halgrímsson, Torsteinn (28. Februar 2003). Web Archiving in Europe [Diskussion].-NWA [online]. – E-mail-Adresse:
nwa@nb.no

[4] Für mehr Informationen zum Paradigma-Projekt siehe unter URL: http://www.nb.no/paradigma/eng. index.html
(Überprüft am15. April 2004)

Methodik und der Geschäftsgänge für die Auswahl, Sammlung, Beschreibung und Identifizierung aller Arten von digitalen Dokumenten – einschließlich jener Dokumente, die im Internet allgemein verfügbar sind. Das Projekt soll den Benutzern den Zugang zu seinem Internet-Archiv unter Einhaltung der gesetzlichen Bestimmungen ermöglichen.

Die Projektaktivitäten bauen auf Vorarbeiten der Nationalbibliothek in verschiedenen relevanten Gebieten auf, und vier Personen sind damit in Vollzeit beschäftigt. Ungefähr dreißig Mitarbeiter beteiligen sich ebenfalls in irgendeiner Form an dem Projekt. Das Projekt soll zum 31. Dezember 2004 abgeschlossen sein.

Die folgenden Abschnitte beschreiben kurzgefasst die laufende Projektarbeit, was das Selektieren, Sammeln und das Zugänglichmachen auf das digitale Pflichtexemplarmaterial anbelangt, sowie die Beschaffenheit und Größe der norwegischen Internet-Domäne.

2.1 Sammel- und Auswahlstrategien

2.1.1 Sammlung

Auf der Basis des Gesetzes zur Pflichtexemplarabgabe und den Empfehlungen des Paradigma-Projekts hat die Nationalbibliothek sich entschieden, mit dem allgemeinen Harvesten aller *allgemein verfügbaren* digitalen Dokumente aus dem norwegischen Web-Bereich („.no“) zu beginnen. Dokumente aus den Bereichen wie z.B. “com“, „.org“ und „.net“ werden ebenfalls eingesammelt.

Es gibt mehrere Gründe dafür, diesen allgemeinen Ansatz für das Harvesting zu wählen: Erstens können wir nicht voraussagen, welche Dokumente später für die Forschung und Dokumentation von Wert sein werden, zweitens wird die digitale Archivierung täglich billiger, drittens erspart ungefiltertes Harvesting die ressourcenintensive manuelle Selektion zum Zeitpunkt des Harvesting, und schließlich kann ein Benutzer des Internet-Archivs Dokumente über Möglichkeiten zur Freitextsuche finden und ist somit in der Lage, alle Dokumente einschließlich derer, für die sich die manuelle Erschließung nicht lohnt, durchsehen. Auswahlkriterien für eine beliebige Weiterverwendung wie z. B. eine weitere bibliografische Beschreibung können jederzeit in Frage gestellt und geändert werden. Dies wäre selbstverständlich unmöglich, wenn das Material beim Harvesting ausgeschlossen worden wäre.

Die Pflichtabgabe-Abteilung hat seit 2001 halb-manuell eine Auswahl von Web-Dokumenten eingesammelt, und zwar unter Anwendung der HTTrack[5]-Software, und diese Dokumente sind im Katalog der Nationalbibliothek (BIBSYS[6]) katalogisiert. Diese Aktivität wird fortgesetzt bis das allgemeine Harvesting des Paradigma-Projekts, und verwandte Verfahren voll etabliert sind. Dieselbe Abteilung sammelt auch an ein Ereignis geknüpfte Dokumente, sie hat z. B. Web-Sites von politischen Parteien vor und nach Wahlen gesammelt. Andere Abteilungen sind ebenfalls an den Aktivitäten zur digitalen Pflichtexemplarabgabe beteiligt, und das Musik- und Bildarchiv der Bibliothek ist damit beschäftigt, Lösungen für die Pflichtexemplarabgabe von originär digitalen Radio- und Fernsehprogrammen in Zusammenarbeit mit der Norwegischen Rundfunkanstalt zu finden.

[5] Für mehr Informationen über die HTTrack-Software siehe unter URL: http://www.httrack.com/ (Überprüft am15. April 2004)
[6] Für mehr Informationen über BIBSYS siehe unter URL: http://www.bibsys.no/english.html (Überprüft am15. April 2004)

Eine äußerst herausfordernde Aufgabe ist die Abgabe des deep web, wie z. B. Internet-Zeitungen, „streaming media“, Dokumente von Web-Kameras, interaktive Medien und elektronische Materialien aller Art, die in Datenbanken gespeichert sind. Das Paradigma-Projekt hat mit der täglichen Sammlung von ca. 65 Internet-Zeitungen begonnen, und es ist geplant, mehrere Zeitungsdatenbanken in naher Zukunft vollständig herunterzuladen und die täglichen „Schnappschüsse“ auf diese Weise zu ergänzen. Wir diskutieren die deep web-Probleme innerhalb des *International Internet Preservation Consortium*[7], aber eine große Anzahl von administrativen, rechtlichen und technischen Fragen sind bis jetzt ungeklärt.

Kurz gefasst erwartet die Nationale Bibliothek von Norwegen, digitale Objekte über mehrere Kanäle zu erhalten: automatisierte Sammlung von Dokumenten aus dem Internet, als Batch gelieferte Datenbank-Updates, Subkriptionszeitschriften und Mailinglisten über e-mail, NetNews-Diskussionsgruppen und Dokumente auf physischen Medien wie den CD-ROMs.

2.1.2 Auswahl

Es gibt im Internet viele nützliche Dokumente, und wir beschäftigen uns gerade damit, *Auswahlkriterien* für diejenigen Dokumente festzulegen, von denen wir glauben, dass sie eine manuelle bibliographische Beschreibung auf irgendeiner Ebene „verdient“ haben. Diese Auswahlkriterien entsprechen dem Gesetz zur Pflichtexemplarabgabe sowie der allgemeinen Sammelpolitik der Bibliothek, wie sie in unserem Leitbild und dem strategischen Plan formuliert ist. Auswahlkriterien für digitale Dokumente werden mit denen für herkömmliche Dokumente im Handbuch der Bibliothek für Dokumentenauswahl integriert.

Das Paradigma-Projekt plant, eine Systemstruktur einzuführen, die eine *Auswahl* in drei Phasen ermöglicht, so dass Bibliothekare technische Hilfe angeboten bekommen, um die wenigen Dokumente zu finden, die auf einer irgendeine Art und Weise katalogisiert werden sollten. In der ersten Phase werden die norwegischen und Sami-Dokumente aus dem Internet aufgefunden und gesammelt. Die zweite Phase gibt den Bibliothekaren die Möglichkeit, automatisch produzierte Ranglisten auf Grund von gezielten Abfragen zu erstellen. Diese Listen basieren auf der Anwendung von Vektoren, die Metadaten enthalten, welche *automatisch* aus den gesammelten Dokumenten extrahiert worden sind. In der dritten Phase suchen Bibliothekare bestimmte Dokumente für die manuelle Erfassung auf irgendeinem Niveau aus den Ranglisten heraus, unter Anwendung der oben genannten Auswahlkriterien. Eines Tages sind wir vielleicht auch in der Lage, „integrating resources“ zu verfolgen, die manuell katalogisiert worden sind und so Bibliothekaren zu helfen, diese bibliographischen Daten in bestimmten zeitlichen Abständen aufzufinden und abzuändern, z. B. wenn Textänderungen einen bestimmten Prozentsatz übersteigen usw.

2.2 Die norwegische Internet-Domain

Die genaue Größe der norwegischen Internet-Domain ist bis jetzt noch immer unbekannt. Die erste Datensammlung seitens des Paradigma-Projekts im Dezember 2002/Januar 2003 resultierte in ungefähr 3,1 Millionen URLs (das heißt Dateien), wovon ungefähr 53 % (durch Zählung) Bilddateien sind (.jpg, .gif, .png). Der NEDLIB-Harvester[8] begann mit zunächst ca. 1.000 URLs, und die Sammlung war begrenzt auf das http-Protokoll, auf die norwegische

[7] Für mehr Informationen über diese tiefe Web-Aktivität siehe unter URL:
http://www.nla.gov.au/ntwkpubs/gw/66/html/p15a01.html (Überprüft am15. April 2004)
[8] Für mehr Informationen über die NEDLIB Harvester siehe URL:
http://www.csc.fi/sovellus/nedlib/ver11/documentation11.doc (Überprüft am15. April 2004)

nationale Domäne („.no“) und auf URLs ohne Parameter. Das zweite Harvesting wurde im August 2003 durchgeführt und resultierte in ungefähr 4,1 Millionen URLs. Die dritte Runde läuft gerade, und bis jetzt gibt es darüber noch keine konkreten Zahlen.

Unter der Annahme einer ähnlichen Verteilung wie beim bereits durchgeführten Harvesting in Schweden und Finnland, erwarten wir, dass 45 % bis 55 % der norwegischen Internet-Adressen in Domains außerhalb von „.no“ gefunden werden. Es versteht sich von selbst – manuelle Behandlung und Auswertung jedes einzelnen Objekts ist nicht möglich; die große Mehrheit muss automatisch bearbeitet werden.

2.3 Zugangsstrategie

2.3.1 *Wer* wird *wonach* in unserem Archiv suchen?

Beim Versuch, für Metadaten Lösungen für die Beschreibung des reichen und vielfältigen digitalen Materials in unserem Archiv zu finden, ist es wichtig zu fragen: Wer wird dieses Material benutzen und zu welchem Zweck? Es ist schwierig, sich die speziellen Fragen von Wissenschaftlern in 10, 20 oder 50 Jahren vorzustellen, aber wir können uns einige *Benutzergruppen* und bestimmte *Arten von Fragen* vorstellen.

Eine Gruppe mag aus Benutzern bestehen, die das Internet und digitales Material als ein Medium studieren wollen, z.B. weil das Material aus dem Internet stammt und es die Charakteristika dieses Mediums zeigt. Hier können wir erwarten, dass einige Benutzer evtl. den Sprachgebrauch im Netz und die Beziehung zwischen unterschiedlichen Sprachformen untersuchen; Medien-Forscher mögen die Beziehung zwischen gedruckten und digitalen Medien oder zwischen technologischen Entwicklungstrends und Inhalt studieren. Benutzer, die das Web-Seiten-Design untersuchen, mögen Interesse an Anzeigen, dem Layout usw. haben. Wissenschaftler auf dem Gebiet der Informatik untersuchen vielleicht unterschiedliche Kommunikationsprotokolle, den Einsatz von Formaten im Lauf der Zeit und vielleicht sogar Viren. Sozialwissenschaftler sind vielleicht daran interessiert, wie die im Internet verfügbaren Informationen die Gesellschaft beeinflusst haben und umgekehrt. Selbstverständlich kann es sich auch um Wissenschaftler mit übergreifenden Interessengebieten handeln.

Eine andere Benutzergruppe mögen jene sein, die digitale Dokumente als Quellenmaterial benutzen müssen – gerade so wie sie heute traditionelle Quellen benutzen. Diese Gruppe wird sicherlich aus Wissenschaftlern aus allen Fachgebieten bestehen, und es ist deshalb interessant herauszufinden, welche Erwartungen sie im besonderen an das digitale Material stellen. Ist das relevante Material allein in digitaler Form verfügbar? Sind dynamischer Inhalt, Animationen, interaktive Displays, integrierte Ton- und Bildwiedergabe usw. von Bedeutung? Brauchen Wissenschaftler den Zugriff über Freitextsuche, oder korrelieren sie große Mengen von Informationen aus verschiedenen Quellen?

2.3.2 Gegenwärtige Gesetzgebung

Es ist eine komplexe Angelegenheit, Benutzern Zugriff auf das im Internet verfügbare Pflichtexemplar-Archiv zu geben, und die Nationalbibliothek muss trotz der vielen, manchmal gegensätzlichen Regelungen in dem Pflichtexemplarabgabe-Gesetz, dem Urheberrecht-Gesetz und dem Personendatenschutz-Gesetz zufriedenstellende Lösungen finden.

Zurzeit versuchen wir, Antworten zu folgenden Fragen zu finden: Welche Benutzer können Zugang zu verschiedenen Arten von digitalen Materialien erhalten? Können sie von Computern außerhalb der Nationalbibliothek auf diese Sammlungen zugreifen?

2.3.3 Zugriffsinstrumente

Die Bedürfnisse der Benutzer, wie zuvor beschrieben, sind für uns von Interesse, da wir versuchen, Zugriffsinstrumente für die Suche in unserem Internet-Archiv zu entwickeln. Wir müssen selbstverständlich die Tatsache berücksichtigen, dass die Bibliothekare wenige der dort verfügbaren Dokumente katalogisieren werden.

Auf einer mehr technischen Ebene hofft man, im Paradigma-Projekt über das Nordic Web Archive's[9] (NWA) Access Tool (siehe Ziffer 1) den Benutzern den Zugriff auf das Internet-Archiv zu geben.

Heutzutage sind die Freitextsuche mit Booleschen Operatoren, die Suche nach einer bestimmten URL und die Präsentation der Dokumentengeschichte über eine Zeitleiste Standardoptionen. Das Zugriffsinstrument wird uns hoffentlich in der Zukunft noch weitere Möglichkeiten bieten: Die Anwendung von Booleschen Suchkombinationen, um verschiedene Trefferlisten zu kombinieren, parallele Suche in katalogisierten Dokumenten in externen bibliographischen Katalogen, Suche in automatisch extrahierten Metadaten, fortgeschrittenes programmiertes Surfen und verfügbare vorprogrammierte Suchparameter, Optionen, die uns Trefferlisten in einer „Projektbibliothek" speichern lassen, Suchzugriff auf Dokumentengruppen, die nach bestimmten Kriterien sortiert sind (Verleger etc.), höchstens ein Treffer für bestehende Dubletten, die Gruppierung eines logischen Dokuments, bestehend aus vielen getrennten Web-Seiten als ein Treffer usw.

Wir planen, die Schnittstelle des NWA Access Tools anzupassen, um verschiedene spezielle Benutzerfunktionen einzupassen, und unsere Anwendung des FRBR-Modells der IFLA wird eine bedeutende Rolle dabei spielen, wie wir in Zukunft den Zugriff auf das archivierte Material gestalten.

3 Die Suche nach Metadaten-Lösungen

Das Paradigma-Projekt befindet sich mitten in seiner Suche nach geeigneten Metadaten-Formaten und –lösungen. Die Definition von Metadaten zum *Auffinden* von Dokumenten war eine unserer Hauptaktivitäten im vergangenen Jahr, ebenso wie unsere Suche nach zufriedenstellenden Lösungen für die automatische Extraktion technischer Metadaten. Im folgenden Abschnitt werden wir versuchen, Ihnen eine kleine Vorstellung davon zu geben, *warum* und *wie* wir beabsichtigen, die vielen digitalen Dokumente in unserem Internet-Archiv zu beschreiben.

3.1 Warum sollten wir Internet-Ressourcen katalogisieren?

Nancy Olsen nennt in der Einführung zu ihrem Buch *Cataloging Internet Resources* drei wesentliche Gründe dafür, *warum* Internet-Ressourcen katalogisiert werden sollten [3]:

1. Es gibt sehr viele wertvolle Informationen, die durch das Internet verfügbar sind.
2. Diese Bestände müssen zwecks Zugänglichkeit organisiert werden.

[9] Für mehr Informationen über das Nordic Web Archive Projekt siehe unter URL: http://nwa.nb.no/ (Überprüft am 15. April 2004)

3. Die Nutzung existierender Bibliothekstechnik und –verfahren und die Erstellung von Datensätzen für das Retrieval in bestehenden Online-Katalogen ist die wirksamste Methode, auf diese Ressourcen zuzugreifen.

Wir stimmen mit Olsen in allen drei Punkten überein, gehen aber gleichzeitig davon aus, dass weit weniger als 1 % des aus der norwegischen Internet-Domain gesammelten Materials jemals irgendwie bibliografisch erfasst wird. Das liegt natürlich an der großen Menge von Dokumenten im Archiv. Wir können versuchen, uns selbst mit folgendem Gedanken zu trösten: Obwohl ein viel höhere Prozentsatz des traditionellen Materials der Bibliothek bibliografisch erfasst wird, werden verschiedene Materialien doch auf verschiedene Art und Weise behandelt: Ephemeres Material wird nur registriert, während man Bücher und Zeitschriften auf einer höheren Katalogisierungsebene erfasst.

Im Gegensatz dazu werden 100 % der Internet-Dokumente mit der FAST[10]-Indexierungssoftware nach dem Harvesting vollständig indexiert. Dies ermöglicht den Bibliotheksmitarbeitern und den Benutzern, im Internet-Archiv sowohl über Freitext als auch über andere Indizes zu suchen. Der winzig kleine Bruchteil von manuell katalogisierten Internet-Dokumenten wird in Volltextform im Archiv und über bibliografische Daten im Bibliothekskatalog verfügbar sein –benutzerfreundlich miteinander verbunden, wie wir hoffen.

Zusätzlich zur Katalogisierung einiger und Indexierung aller Dokumente werden wir vorhandene eingebettete Metadaten sowie die Internet-Dokumente, die sie beschreiben, einsammeln. Die Nationalbibliothek plant für die Zukunft einen Service, der es den Verlegern ermöglicht, Metadaten zu erzeugen und bei der Abgabe ihrer Dokumente zu liefern.

3.2 Was sind Metadaten?

Die Suche nach Metadatenlösungen hat natürlich auch zur Suche nach angemessenen Definitionen geführt. Der Begriff „Metadaten“ wurde in der Literatur immer wieder neu definiert. „Daten über Daten“ ist vielleicht die am häufigsten wiederkehrende Definition, und Metadaten umfassen einen ganzen Bereich von Informationsarten[11]. Wir haben entdeckt, dass Metadatenschemata so vielfältig wie unterschiedlich sind, aber sie haben eines gemeinsam: Sie können uns helfen, die vielen nützlichen Dokumente in unserer Sammlung zu beschreiben und zu finden – auch jene, die keine „Kandidaten“ für die Katalogisierung auf hohem Niveau sind.

3.3 Was ist ein Internet-Dokument?

3.3.1 Definition eines Internet-Dokuments aus technischer Sicht

Wenn ein Internet-Dokument zum Harvesting und deshalb zum Archivieren ausgewählt worden ist, kann die Bedeutung des Begriffs „ein Dokument“ höchst vieldeutig sein: Welche Bestandteile sollten eingesammelt und als feste Bestandteile des Dokumentes archiviert werden? Welche Komponenten sollten der individuellen Bewertung unterliegen? Wir setzen voraus, dass jeder Bestandteil, der das „Aussehen“ einer Web-Seite beeinflusst (einschl. *Ton*

[10] Für mehr Informationen über FAST Search & Transfer (FAST) ASA siehe unter URL: http://www.fast.no (Überprüft am 15. April 2004)

[11] *Einer von den vielen Metadaten-Berichten, die wiruns angesehen haben, ist: DESIRE*: A review of metadata: a survey of current resource *description formats* (1997). Siehe URL: http://www.ukoln.ac.uk/metadata/desire/overview/rev toc.htm (Überprüft am 15. April 2004)

und andere *nicht grafische* Elemente), unbedingt aufgenommen werden sollte, wenn eine Web-Seite ausgewählt wird, d.h. Hintergrundabbildungen, Rahmeninhalte, Abbildungen für Buttons usw.

Durch Links referenzierte Dokumente sind unterschieden von, sind jedoch bezogen auf, das referenzierende Dokument. Auf einem höheren semantischen Niveau wollen wir oft eine ganze Gruppe von Dokumenten, die miteinander verknüpft sind, als ein einziges großes Dokument behandeln. Wenn wir sie als ganz unabhängige Dokumente behandeln, gehen wir das Risiko ein, beispielsweise wenige Kapitel aus einem Bericht herauszupicken, während wir andere Kapitel auslassen (dies wäre möglich, weil sie breite Zitate, Zusammenfassungen usw. in anderen Sprachen als Norwegisch enthalten).

In Beantwortung unserer Frage „Was umfasst ein Internet-Dokument?“ können wir sagen, dass ein Internet-Dokument aus vielen verknüpften Teilen oder Dateien besteht, z.B. Text, Abbildung, Ton, Animation, usw. und diese meistens durch Links verknüpft und manchmal in Frame Sets enthalten sind.

3.3.2 Definition eines Internet-Dokuments aus bibliografischer Sicht

Wir können uns selbstverständlich niemals auf einen Computer verlassen, der uns sagen soll, wo ein Internet-Dokument anfängt und wo es endet – selbst wenn wir ihn mit diesem Ziel so programmieren, dass er bestimmten Anweisungen folgt. Glücklicherweise können Bibliothekare sehr gut entscheiden, welche der vielen Teile eines Internet-Dokuments ein logisches Ganzes ergeben. So können wir aus bibliografischer Sicht ein Internet-Dokument als eine Informationseinheit definieren, die bibliografisch beschrieben werden kann. Diese Definition bestimmt *nicht* von vorneherein feste oder eindeutige Dokumentbestandteile, sondern lässt stattdessen den Bibliothekar das zu beschreibende Objekt bestimmen: Eine gesamte Web-Site kann durch einen Datensatz beschrieben werden, und ein bestimmtes Dokument auf dieser Site kann man ebenfalls eine Beschreibung erhaltenn. Der Bibliothekar kann Hintergrundgeräusche, Style Sheets usw. einbeziehen oder weglassen, und er kann mehrere eng aufeinander bezogene Web-Seiten, z.B. Kapitel eines Berichts, in ein Dokument zusammenfügen. In der Zukunft werden unsere automatisierten Verfahren dem Bibliothekar Dokumentdefinitionen vorschlagen, basierend auf einer Analyse des Inhalts, Verknüpfungsgruppen etc.: Durch eine Standardeinstellung werden eingebettete Abbildungen, direkt referenzierte stehende Musik-/Videoclips und Style Sheets in das Dokument aufgenommen. Links einer bestimmter Art, die eine referenzierte Web-Seite identifizieren, z. B. als Inhaltsverzeichnis oder als ein Abschnitt, werden ebenfalls eingefügt.

Somit ist eine Informationseinheit, die bibliografisch beschrieben werden kann, der Ausgangspunkt für eine Beschreibung mit Metadaten, sowohl wenn digitales Material auf festen Trägern wie den CD-ROMs, DVDs gespeichert ist, als auch wenn es als separate Dateien aus dem Internet gesammelt wird. Das bedeutet, dass alle digitalen Dokumente – angefangen bei den *traditionellen* Dokumenten wie Monografien, Dissertationen etc., den *flüchtigen* Dokumenten wie den Internet-Zeitungen, Hyper-Poetry, Hyper-Drama usw. und schließlich den *neuartigen* Dokumenttypen wie Homepages, Web-Logs (d.h. blogs) usw. Kandidaten für eine Beschreibung mit Metadaten innerhalb unseres Internet-Archivsystems sind.

3.4 Metadaten-Bestandsaufnahme und damit verbundene Arbeit

3.4.1 Welche Arten von Metadaten brauchen wir?

Wir fanden die Frage interessant, welche Metadatenformate die Nationalbibliothek heute für die Beschreibung verschiedener Arten digitalen Materials benutzt. Diese Information kann nützlich sein, da wir hoffen, eines Tages in der Lage zu sein, Daten in unser Archiv zu importieren und zu exportieren. Die Ergebnisse unserer Übersicht zeigen, dass verschiedene Formate benutzt werden: BIBSYS-MARC (das MARC-Format des BIBSYS-Systems) für digitalen Text, Dublin Core Metadata Element Set[12] für Radioprogramme, MAVIS[13] (ein australisches System und Format) für Fernsehprogramme, Ton und Bilder sowie andere Formate, die in lokal entwickelten Systemen angewendet werden.

Die Metadatenformate sind für ihre Anwendung sehr gut geeignet, aber sie sind keine zufriedenstellende Lösungen für all unsere Metadatenanforderungen. Das Internetarchiv benötigt mehrere Arten von Metadaten: *administrative Metadaten,* z.B. für die Erstellung und Änderung von Metadatensätzen; Metadaten für die *Rechte- und Zugriffsverwaltung*, um Urheberrechtsinformationen zu speichern und festzulegen, welche Benutzergruppen Zugriff auf das Archiv erhalten und welche Dokumente sie lesen können; *strukturelle* Metadaten für die Erläuterung der logischen Beziehungen zwischen Objekten, zwischen Metadaten oder zwischen Objekten und Metadaten; *Metadaten zur langfristigen Speicherung* für die Festlegung von z. B. Dateiarten, notwendige Software und Dokumentkonversion/Migrationsverlauf, und schließlich *technische* Metadaten zur Spezifizierung von Dokumentengröße, Skripten, Datenübertragungsdetails usw. Zu guter Letzt brauchen wir *beschreibende* und *analytische* Metadaten für Such- und Retrievalzwecke.

3.4.2 Welches Beschreibungsmodell sollten wir wählen?

Es gibt unterschiedliche Ansichten darüber, welches Beschreibungsniveau ein digitales Dokument erhalten sollte. Bei unserer Arbeit, Metadaten für beschreibende und analytische Metadaten zu definieren, betrachteten wir zwei alternative Modelle. Die eine Alternative ist, drei Beschreibungsstufen anzuwenden:

1. Katalogisierung für die Einbeziehung in die Nationalbibliografie / den Katalog BIBSYS der Nationalbibliothek / andere spezielle Datenbanken.
2. Katalogisierung auf einer einfacheren Stufe in einem verbreiteten Format.
3. Automatische Extraktion von Metadaten sowohl aus dem Dokument selbst als auch aus den Datenübertragungsprotokollen usw.

Die andere Alternative ist die Anwendung eines Zweistufen-Modells, d.h. „zu katalogisieren – oder nicht zu katalogisieren“:

1. Katalogisierung für die Einbeziehung in die Nationalbibliografie/den Katalog BIBSYS der Nationalbibliothek / andere spezielle Datenbanken
2. Automatische Extraktion von Metadaten sowohl aus dem Dokument selbst als auch aus den Datenübertragungsprotokollen usw.

Es gibt mehrere Argumente für diese zweite Alternative: 1) Das Retrieval von digitalem Material (Freitext etc.) ist nicht abhängig von der Erfassung, wie es bei nicht erfasstem Analogmaterial der Fall ist. 2) Es ist für die Bibliothek nicht nötig, Material zu erfassen, um

[12] Für mehr Informationen über die Dublin Core Metadata Initiative siehe unter URL: http://www.dublincore.org) (Überprüft am am 15. April 2004)
[13] Für mehr Informationen über Wizard’s MAVIS system siehe unter URL: http://www.wizardis.zoom.au/ie4/products/mavis/introducingmavis.html (Überprüft am 15. April 2004)

seine Verbreitung verfolgen zu können, z. B. welche Universitätsbibliotheken Kopien erhalten haben. 3) Wir können jederzeit unsere Entscheidung bedauern, eine bestimmte Art von digitalem Material nicht zu katalogisieren.

Eine kurze Darstellung jeder einzelnen der drei Stufen wird im folgenden Abschnitt gegeben.

∅Katalogisierung zur Aufnahme in die Nationalbibliografie usw.

Zur Zeit sind unsere Vorschläge, welche Arten von Dokumenten auf diesem höchsten Niveau katalogisiert werden sollten, noch unvollständig, aber wir können mit Sicherheit sagen, dass eine kleine Anzahl von nützlichen digitalen Dokumenten weiterhin in einem MARC-Format für die Aufnahme in die Nationalbibliografie katalogisiert werden. (Wir möchten erwähnen, dass Norwegens MARC-Version NORMARC heißt, dass einige Systeme lokale Versionen übernommen haben, z. B. BIBSYS MARC, und dass die Anwendung von MARC21[14] zurzeit auf nationaler Ebene diskutiert wird. Norwegens Katalogisierungsregelwerk basiert auf der zweiten Ausgabe der Anglo-American Cataloguing Rules (AACR2), und die Kapitel 9 und 12 sind jetzt in Norwegisch verfügbar.)

Wir können außerdem mit Sicherheit sagen, dass die Katalogisierung audiovisuellen Materials für die langfristige Erhaltung ein hohes Maß an Details erfordert – besonders wenn es darum geht, die Übersicht über die technischen Informationen hinsichtlich Wiederherstellung von Originalen, Kopien usw. zu behalten. Die Bibliothek wird zweifellos weiterhin MAVIS für diese Arbeit benutzen.

∅Katalogisierung auf einer einfacheren Ebene in einem verbreiteten Format

Wie zuvor erwähnt plant die Nationalbibliothek zukünftig einen Service, der es Verlegern erlaubt, Metadaten zu erzeugen und zusammen mit den Dokumenten bei der Abgabe zu liefern. Das Paradigma-Projekt arbeitet gegenwärtig daran, um das (die) Metadatenformat(e) zu definieren, welche(s) in Zukunft die Grundlage für ein benutzerfreundliches Werkzeug, bereitgestellt durch diesen Service, bilden wird. Möglicherweise werden Bibliothekare einmal die von Verlegern gelieferten Metadaten vielleicht als Grundlage für übergeordnete bibliographische Daten benutzen können.

Wir haben einige Metadatenbankformate analysiert und verglichen, um geeignete Lösungen zu finden:
Machine Readable Cataloguing (MARC) und Dublin Core Metadata Element Set (DCMES), da diese beiden in Bibliotheken und ähnlichen Institutionen benutzt werden; Metadata Object Description Schema (MODS)[15] und Metadata Encoding & Transmission Standard (METS)[16], da diese von Bibliotheken für Bibliotheken entwickelt wurden, und Online Information eXchange (ONIX)[17], da dieses Format von der Verlags- und Buchindustrie entwickelt worden ist. Wir möchten auch festhalten, dass die ISBN-Gemeinschaft vorgeschlagen hat, dass sich

[14] Für mehr Informationen über MARC21 siehe unter URL: http://www.loc.gov/marc/bibliographic /ecbdhome.html (Überprüft am ab 15. April 2004)
[15] Für mehr Informationen über MODS siehe unter URL: http://www.loc.gov/standards/mods/ (Überprüft am 15. April 2004)
[16] Für mehr Informationen über METS siehe unter URL: http://www.loc.gov/standards/mets/ (Überprüft am 15. April 2004)
[17] Für mehr Informationen über ONIX siehe unter URL: http://www.loc.gov/standards/mets/ (Überprüft am 15. April 2004)

registrierende Einrichtungne die ISBN-Agenturen mit ONIX-kompatiblen Metadaten in Verbindung mit der Zuteilung einzelner ISBN versorgen können.

Wir haben die obigen Formate anhand folgender Fragen verglichen: Wer ist für die Verwaltung des Formats verantwortlich? Ist es ein internationaler Standard? Für welches Gebiet wird es benutzt? Welche Art von Medien beschreibt es? Schließt es semantische und/oder syntaktische Definitionen ein? Wie beschreibt es Verknüpfungen von Dokumenten untereinander? Ist das Format von bestimmten Regeln oder Codierungen abhängig? Ist es kompatibel mit oder steht es in Beziehung zu anderen Formaten? Wie umfangreich wird es angewendet und von welchen Gemeinschaften?

Wir hoffen, dass diese Bestandsaufnahme in der Bibliothek zu einer breiteren Diskussion über Metadaten in Verbindung mit einer fortlaufenden Aktualisierung führen kann. Wir haben ebenfalls vor, näher zu untersuchen, inwieweit Benutzeranforderungen durch den Gebrauch der *Common core records*, wie sie durch die IFLA-Arbeitsgruppe für die Anwendung von Metadatensystemen [6] und IFLA's *Final Report on Functional Requirements for Bibiographic Records* (FRBR) [5] vorgeschlagen werden, erfüllt werden können. Die Zusammenarbeit auf dem Gebiet von Metadatenlösungen mit einem laufenden bibliografischen Projekt innerhalb der Nationalbibliothek sowie der *Norwegian Digital Library*, einem anderen Projekt auf nationaler Ebene, stehen ebenfalls auf unserem Programm. Wir hoffen, dass diese Arbeit in die Empfehlung von Metadatenformaten für die Beschreibung auf verschiedenen Ebenen resultiert.

In der Zwischenzeit haben wir daran gearbeitet, die Anforderungne an die technischen Metadatenfür unserer Archivsystemsoftware zu spezifizieren. Wir haben verschiedene Faktoren ermittelt, die unsere Wahl eines Metadatenformats für die Katalogisierung auf unterem Niveau beeinflussen können. Hier sind einige technisch wünschenswerte Faktoren:

- Semantische Interoperabilität mit MARC: Es ist wichtig, dass die Attribute des Metadatenformats mit dem in Bibliotheken eingesetzten MARC-Format semantisch harmonisieren. Wenn möglich sollte das Format ein funktioneller Ausschnitt von MARC sein. Dies würde den Datentausch vereinfachen.

- Einfach aber inhaltsreich: Es ist wichtig, ein Metadatenformat zu finden, welches einfach zu benutzen, jedoch inhaltsreich genug ist, eine angemessene Menge von Einzelheiten zu repräsentieren.

- Einfach in andere Formate konvertierbar: Eine Konvertierung zwischen dem gewählten Format und MARC sollte verfügbar oder relativ einfach festzulegen sein. Wir stellen fest, dass Konkordanzen MARC21 und MODS, zwischen MARC21 und ONIX und zwischen Dublin Core ohne Qualifikatoren und MODS bereits existieren.

- XML-Kompatibilität: XML ist mehr oder weniger ein Defacto-Standard, und ein Format, das XML-kompatibel ist, wird uns die Bearbeitung des Formats mit der verfügbaren Software erlauben. Eine umfassende Rahmenstruktur wird ebenfalls in XML festgelegt werden, so dass Metadaten aus verschiedenen Quellen vom Archiv aufgenommen, Änderungen von Metadaten bearbeitet, Original-Metadaten bestimmt und die Versionsgeschichte verfolgt werden können usw. (z.B. METS).

- Erweiterungsmöglichkeit: Ein Metadatenformat sollte wenn nötig die Festlegung neuer Elemente erlauben.

- Kernelemente: Es ist wichtig, Metadatenkernelemente festzulegen, z. B. eine allgemeine Bezeichnung, die die Dokumentensuche und die Abfrage unter unterschiedlichen Arten von Material erleichtern kann.

Wenn wir diese Faktoren mit den Metadatenformaten in unserer Bestandsaufnahme vergleichen, sehen wir, dass den Formaten, die MARC- und XML-kompatibel sind, der Vorzug zu geben ist. Es gibt jedoch kein einfaches Rezept dafür. Es müssen neue Elemente für technische, strukturelle, auf Rechte und Zugriff bezogene Metadaten festgelegt und evtl. innerhalb des METS-Systems zusammengeführt werden. Das Gleiche gilt selbstverständlich auch für Metadaten für den langfristigen Erhalt. Hier verlangt der Langzeitspeicheer der Bibliothek die Anwendung von OAIS[18]-konformen Metadaten.

∅Automatische Extraktion von Metadaten

Leider werden Bibliothekare niemals überwältigende 99 % der Internet-Dokumente in unserem Archiv katalogisieren. Deshalb prüfen wir gerade die Möglichkeit einer automatischen Analyse und Extraktion von Metadaten aus Internet-Dokumenten als Teil unserer Arbeit mit Metadaten und Systemdesign. Extrahierte Metadaten werden zusammen mit den digitalen Objekten und anderen Metadatenbeschreibungen aufbewahrt und für strukturiertes Suchen im Internet-Archiv verfügbar gemacht.

Die Technologie ist bis jetzt nicht gut genug, um automatisch einen Dokumententyp festzulegen, aber sie kann helfen, die Anzahl von Dokumenten zu verringern, denen man sich in Phase 2 unseres Auswahlprozesses widmen muss. Beispiele für solche Eigenschaften von Dokumententypen sind 1) Sprache, Vokabular und Grammatik; 2) Dokumentgröße und –struktur; 3) Quelle/Verleger/Web-Server; 4) Gebrauch von „cookies“; 5) Alters- und Lebenserwartung eines Dokuments; 6) Ton, Bilder, Animationen, Video und andere fortgeschrittene Informationsarten; 7) Interaktion mit dem Benutzer durch „Eingabemasken“, Buttons etc.; 8) Anzahl, Art und Quelle der Links; 9) URL-Werte, z. B. Gebrauch von bestimmten Worten oder Buchstaben in der URL; 10) Nutzung kundenseitiger Skripte; 11) Einzelheiten der technischen Datenübertragung.

Die Technologie zur Analyse des Vokabulars und der Grammatik verbessert sich, und wir glauben, dass diese Analyseart in der Zukunft ein wichtiger Bestandteil der automatischen Arbeitsverfahren sein kann. Möglicherweise einmal werden die automatisch ausgewählten Typ-Eigenschaften für das strukturierte Suchen im Internet-Archiv zur Verfügung stehen. Der Nutzen dieser Eigenschaften wird begrenzt sein, aber in Kombination mit anderen Suchkriterien mögen sie sich doch als nützlich erweisen.

4 Die Rolle der FRBR im Internet-Archiv

Das Paradigma-Projekt möchte die archivierten digitalen Dokumente und Metadaten in einer organisierten und strukturierten Weise präsentieren, um die Benutzernavigation zu vereinfachen. Wir haben festgestellt, dass das FRBR-Modell der IFLA dabei ein wesentliches Instrument ist, und wir werden dieses Modell als Grundlage für das Design des Internet-Archivs benutzen.

[18] Für mehr Informationen über das OAIS Reference Model siehe unter URL: http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf (Überprüft am 15. April 2004)

Wir glauben, dass die Hinzufügung von Mechanismen für Aggregatmodellierung zum FRBR-Modell unsere Arbeit mit dynamischen Medien wie z. B. Internet-Dokumenten, Multimedia und anderen fortlaufenden Ressourcen fördern wird. Aggregatmechanismen können als einfache Erweiterungen des Modells implementiert werden, ohne größere Änderungen in der existierenden FRBR-Konzeption erforderlich zu machen. Ein Aufsatz über die von uns vorgeschlagenen Aggregatmechanismen wird in diesem Jahr im FRBR-Themenheft von *Cataloging & Classification Quarterly* veröffentlicht.

Um das FRBR-Modell an dynamische Internet-Dokumente anzupassen, wird eine moderate Neuinterpretation der Konzepte für *manifestation* und *item* benötigt, sie wird im folgenden Abschnitt beschrieben.

4.1 Anpassungen der FRBR für die Anwendung auf dynamische Internet-Dokumente

4.1.1 Dynamische Dokumente

Internet-Dokumente sind oft dynamisch, z.B. eine Internet-Zeitung, die mehrmals am Tag aktualisiert wird. Ein Benutzer mag diese Art von dynamischen Dokumenten als ein Forum oder Informationskanal verstehen: „The Daily News berichten, dass ...“. Man kann vielleicht sagen, dass ein dynamisches Dokument ungefähr einer URL entspricht. Konzepte von „Heften“ und fortlaufenden „Ausgaben“ müssen im Zusammenhang mit dem Internet ebenfalls überdacht werden: Aus formaler Sicht mag die Aktualisierung einer Web-Seite einer neuen Buchauflage ähnlich sein. Jedoch betrachten Leser z.B. die ständig sich verändernde Titelseite einer Internet-Zeitung als eine einzige sich verändernde Einheit – nicht als gesonderte, separate Ausgaben.

Unter Anwendung des FRBR-Modells mit Erweiterungen für Aggregatkomponenten haben wir den Begriff *dynamisches Dokument* als „den gesamten Lebenszyklus einer ständig sich verändernden Web-Seite oder eines ähnlichen Internet-Dokuments“ definiert.

Wenn wir ein sich aktualisierendes Web-Dokument dieser Art gemäß AACR2 zu katalogisieren hätten, würden wir normalerweise die Regeln für Integrating Resources anwenden, d. h. eine bibliografische Ressource, die erweitert wird um oder verändert wird durch Updates, die nicht für sich alleine stehen und in das Gesamtdokument integriert werden. Jedoch ähneln Dokumente wie Internet-Zeitungen eher Radioprogrammen, einem ständig wechselnden Fluss von flüchtigen Informationen. Sie „ordnen sich nicht als das Ganze zusammen“. Den Inhalt eines sich ständig verändernden Dokuments zu einem festgelegten Zeitpunkt zu erfassen ist wie die Aufzeichnung einer *Probe* einer flüchtigen Rundfunksendung. Wir bezeichnen jede dieser Proben oder Momentaufnahmen als *spezifisches Dokument*.

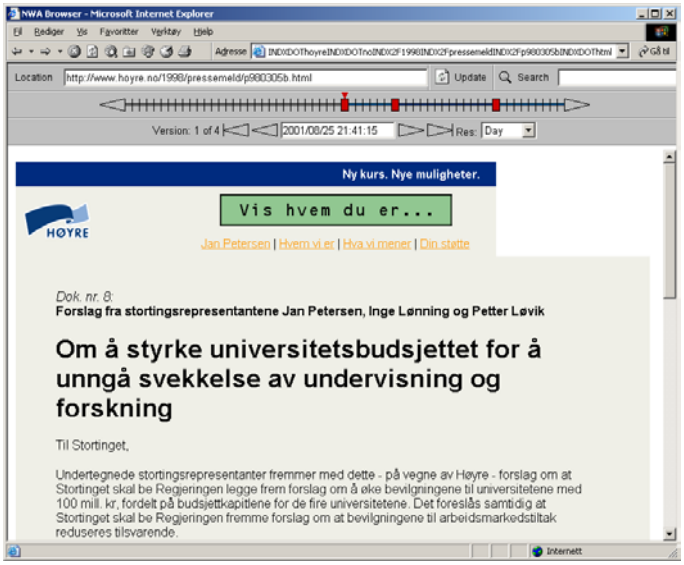
Wenn auf ein dynamisches Dokument im Internet zugegriffen wird, kann sich das *item* (d. h. die konkrete Veranschaulichung), das von einem Benutzer abgefragt wird, von allen anderen *items* desselben Dokuments unterscheiden: Es kann von der Kombination einer Anzahl von Faktoren abhängig sein: Benutzeridentität, das benutzte Zugangsinstrument (Web-Browser), Informationen über frühere Zugriffe auf dasselbe Dokument (gespeichert in Cookies), Parameter, die ausdrücklich vom Benutzer z. B. in einem Formular festgelegt werden, und zu guter Letzt der aktuelle Stand einer Datenbank. Oft wird das *item* on the fly erzeugt, dann, wenn ein Benutzer eine Darstellung anfordert. Mit anderen Worten, ein http-Abruf funktioniert wie ein Print-on-Demand-Service: Die gelieferte Kopie spiegelt wider, was auch immer der Inhalt der Dokumentendatenbank zum Zeitpunkt des Druckens ist. Die Datenbank

kann als eine (semi-)permanente physische Darstellung des dynamischen Dokumentes angesehen werden, von der bestimmte *items* abgeleitet werden können. Die *items* selbst haben keine dauerhafte Repräsentation, sie sind flüchtig, es sei denn, sie werden z. B. in einem Internet-Archiv gespeichert.

4.1.2 Spezifische Dokumente

Wir haben ein *item*, das ein dynamisches Dokument exemplifiziert, als ein *spezifisches Dokument* definiert, das von einem herkömmlichen *item* vor allem in einer Hinsicht abweicht: Es gehört zu einer Gruppe von *items*, welche dasselbe dynamische Dokument exemplifiziert. Ein Dokument, das im Archiv gespeichert oder dem Benutzer auf dem Bildschirm angezeigt wird, ist offensichtlich ein bestimmtes Dokument, aber dies wird abgeschwächt: Eine Volltextsuche wird höchstens einen einzigen Eintrag in der Trefferliste für ein dynamisches Dokument aufweisen. Wenn ein Benutzer sich den Treffer anzeigen lassen möchte, wird das dynamische Dokument als eine Einheit präsentiert, und der Benutzer kann dann ein bestimmtes *Item* auf einer *Zeitleiste* auswählen, z.B. einer *Menüleiste* (Befehlsleiste), in der die Lebensspanne des Dokuments dargestellt wird. Jede gespeicherte Version, d. h. jedes spezifische Dokument, wird in der Zeitleiste mit einem Markierungszeichen angezeigt. Der Benutzer kann auf jedes spezifische Dokument zugreifen, indem er den Marker für ein(e) bestimmte(s) Datum/Zeit anklickt, um so das *item* abzufragen (siehe Ziffer 1).

Abbildung 1. Präsentation eines dynamischen Dokuments über die Benutzerschnittstelle des NWA Zugangs-Tools



4.2 Von Verlagen oder Benutzern festgelegte Definitionen für Dokumente und Metadaten

Die Präsentation archivierter Dokumente für wissenschaftliche und dokumentarische Zwecke ist nur eine Dienstleistung, die von der Nationalbibliothek bereitgestellt werden wird. Zusätzlich und basierend auf den obigen Überlegungen haben wir Revisionen für den bestehenden ‚Service der Bibliothek für die Zuteilung von Identifiern vorgeschlagen. Derzeit teilt dieser Web-Service den Universitäten und anderen Institutionen die URN:NBNs [7] aus

dem norwegischen Bereich des URN:NBN-Namensbereichs zu. Wir sehen allerdings die Möglichkeit, einzelne ISBNs von diesem Service ebenfalls zuteilen zu lassen.

4.2.1 Zukünftige Funktionalität – ein Szenario

Ein Szenario, das die zukünftige Funktionalität zeigt, ist wie folgt: Die erste Serie von Identifiier, die von diesem Service zugeteilt wird, verlangt vom Benutzer/der anfordernden Stelle die Lieferung eines Mindestsatzes an Metadaten und eine exakte Beschreibung des identifizierten Dokuments.

Es können Identifier für *work, expression, manifestation* (einschließlich Festlegungen dynamischer Dokumente) und *item* (einschließlich Festlegungen spezifischer Dokumente) zugeteilt werden. *Items* (spezifische Dokumente) müssen durch eine vollständige Liste von Bestandteilen (z. B. eine HTML-Datei, Bilderdateien, Musikdateien usw.) spezifiziert sein; *manifestations* (dynamische Dokumente) können ebenfalls durch Regeln festgelegt werden, solche wie „Die Titelseite der Internet-Zeitung in dieser URL und alle Seiten, direkt von der Titelseite verknüpft, welche sich auf derselben Site befinden".

Für die Identifier für *expression* und *work* kann der Benutzer zwischen den Optionen *expressions*/dynamische Dokumente und *items*/spezifische Dokumente wählen, die Umschreibungen dieses *work*/dieser *expression* sind.

Verlags- oder benutzerseitig definierte Beschreibungen werden eher als die automatisch vorgeschlagenen als endgültig angesehen. Die Identität des Verlages oder Benutzers, welche den Identifier zuweist, wird archiviert; die Definition eines Dokuments, welche durch ein anerkanntes Verlagshaus oder eine Universität vorgenommen wird, kann als bedeutender als die von einem beliebigen Benutzer angeforderte Definition eingeschätzt werden.

4.2.2 Metadatenfelder

Obligatorische und nicht obligatorische Metadatenfelder könnten für die Beschreibung des Dokuments auf jeder FRBR-Stufe verfügbar gemacht werden, und jede Stufe würde mit einer URN:NBN gekennzeichnet werden. Die Metadatenwerte werden mit dem Identifier gespeichert, und Benutzer, die unseren Internet-basierten Auflösungsservice nutzen, werden in der Lage sein, das Dokument unter dieser Nummer zu finden.

Nach Eingabe der Informationen in die Metadatenfelder eines zukünftigen Metadaten/Zuteilungstools könnte ein Verlagsmitarbeiter auf einen Button klicken, z. B. <HTML Dublin Core>, um sich diese Metadaten dann in der HTML-Version in einem separaten Fenster anzusehen. Der Benutzer könnte diese Metadaten anschließend kopieren und sie in das <HEAD>-Element des beschriebenen Web-Dokuments einfügen, bevor die Zuteilung der Identifier fortgesetzt wird. Nach Sicherung des digitalen Dokuments, das jetzt eingebettete Metadaten enthält, kann der Benutzer leicht die durch Metadaten angereicherte Dokumentenkopie im Bibliotheksarchiv durch Betätigen des Update-Knopfes des Browsers ablegen.

4.3 Geeignete Programme zur Nachweis- und Authentizitätsprüfung von Dokumenten?

Wir haben Geschichten über Behörden gehört, die offizielle Stellungnahmen im Internet überarbeiteten und sich später geweigert haben, die Existenz früherer Versionen zu bestätigen.

Wir haben auch von kommerziellen Unternehmen gehört, die ihre Produkte zu einem bestimmten Preis anbieten und den Kunden dann mit einem viel höheren Betrag belasten.

Um dem vorzubeugen schlägt das Paradigma-Projekt einen Verifizierungs- und Authentifizierungs-Service vor, mit dem die Benutzer den Download eines bestimmten Internet-Dokuments anfordern können, z. B. eine Momentaufnahme einer Web-Seite, die ein besonderes kommerzielles Angebot, Angaben über rechtliche Verantwortlichkeit, Verleumdungen usw. enthalten.

Sollten hinsichtlich dieser Dokumenteninhalte zu einem späteren Zeitpunkt Zweifel auftreten, könnte die Bibliothek Forderungen in dieser Hinsicht bestätigen (oder ablehnen). Selbst wenn keine rechtlichen Aspekte vorliegen, kann das gespeicherte *item* eines bestimmten Dokuments als ein wohldefiniertes Abbild eines dynamischen Dokuments zu einem bestimmten Zeitpunkt dienen, z. B. zum Zitieren und Verweisen. Dies ist wichtig, wenn wir bedenken, dass die meisten Internet-Dokumente keine Seitenzahlen, keine Versionsnummern etc. besitzen.

In unserem Internet-Archiv wird ein spezifisches Dokument in der Form definiert, in der es vom Web-Server empfangen wurde. Es gibt einen wohldefinierten Bit-Strom für jeden Bestandteil des Dokuments (Text, Bilder usw.). Die graphische Wiedergabe des Dokuments ist *nicht* Teil seiner Definition – dieser Prozess bleibt dem Zugangstool überlassen. Das spezifische Dokument wird durch bestimmte Bestandteile und Metadaten als Inhalt eines dynamischen Dokuments identifiziert:

- die Quelle jeder einzelnen Komponente (z. B. eine URL)
- alle vom Kunden spezifizierten Parameter bei der Abfrage der Bestandteile
- die Istzeit der Abfrage jeder einzelnen Komponente
- die Liste von Bestandteilen, die in dem Dokument enthalten sind

5 Schlussfolgerung

Das Paradigma-Projekt der Nationalbibliothek in Norwegen arbeitet intensiv daran, eine zufriedenstellende Technologie, Methodik und Geschäftsgänge für die Pflichtexemplarabgabe aller Arten von digitalen Dokumenten – auch der Millionen von Dokumenten, die in der norwegischen Internet-Domäne gefunden wurden, innerhalb der verbleibenden Projektzeit auszuarbeiten. Wir hoffen, unseren Benutzern den Zugriff auf das Archivmaterial über bibliografische Datensätze, diverse Arten von Metadaten- und Instrumente zur Volltextsuche bereits in 2005 zu ermöglichen.

Unser FRBR-strukturiertes Internet-Archiv wird sicherlich eines der ersten seiner Art sein, und wir hoffen, dass wir unsere Ideen für die Anwendung der FRBR-Entitäten *work, expressioin, manifestation* und *item* in Zukunft in einer Dienstleistung zur Vergabe von Identifiern realisieren können. Vielleicht werden unsere Vorstellungen von einem Internet-Service zum Nachweis/zur Authentizitätsprüfung irgendwann in der Zukunft auch Wirklichkeit? Die Zeit wird es zeigen, aber in der Zwischenzeit wird die Nationalbibliothek weitermachen mit der Erkundung neuer Wege zum Erhalt des digitalen Kulturerbes Norwegens und der Bereitstellung von Instrumenten für die Benutzer, mit denen sich die Türen zu dieser aufregenden digitalen Bibliothek öffnen lassen.

Ausgewählte Referenzen

(Alle URLs wurden am 15. April 2004 überprüft.)

[1] Abschlussbericht für das Pilotprojekt „Netarkivet.dk“ [online]. -
URL: http://www.netarkivet.dk/rap/webark-final-rapport-2003.pdf

[2] Guidelines for the selection of online Australian publications intended for preservation by the National Library of Australia [online]. –
URL: http://pandora.nla.gov.au/selectionguidelines.html

[3] The Kulturarw3 Project – The Royal Swedish Web Archiw3e – An example of “complete” collection of web pages [online]. –
URL: http://www.ifla.org/IV/ifla66/papers/154-157e.htm

[4] Olsen, Nancy (2002). Cataloging Internet Resources : A Manual and Practical Guide [online]. OCLC. –
URL: http://www.oclc.org/support/documentation/worldcat/cataloging/internetguide/1/1.htm

[5] Norwegen. [Das Gesetz zur Pflichtexemplarabgabe (1989)]. Gesetz zur Pflichtexemplarabgabe von allgemein verfügbaren Dokumenten: Nr. 32 vom 9. Juni 1989: mit Regelungen / [veröffentlicht durch das Ministerium für Kirche und Kulturelle Angelegenheiten; inoffizielle englische Übersetzung, veröffentlicht durch die Nationalbibliothek von Norwegen. - [Oslo]: Nationalbibliothek von Norwegen, 1997. – 21 S.

[6] IFLA Cataloguing Section Working Group on the Use of Metadata Schemas (2003). Guidance on the structure, content, and application of metadata records for digital resources and collections : draft for worldwide review 27 October, 2003 [online]. – URL:
http://www.ifla.org/VII/s13/guide/metaguide03.pdf

[7] RFC 3188 Using National Bibliography Numbers as Uniform Resource Names [online] / J. Hakala, 2001. – URL: http://www.ietf.org/rfc/rfc3188.txt

[8] Van Nuys, Carol (2003). Identification of network accessible documents : problem areas and suggested solutions [online] / Carol van Nuys, Ketil Albertsen. – S. 13-25. – *In*: Proceedings : in conjunction with the 7th European Conference on Research and Advanced Technologies for Digital Libraries, ECDL 2003 / Julien Masanès, Andreas Rauber, Gregory Cobena (eds). – URL: http://bibnum.bnf.fr/ecdl/2003/index.html

[9] Albertsen, Ketil (2003). The Paradigma web harvesting environment. – S. 49-62. – *In*: Proceedings : in conjunction with the 7th European Conference on Research and Advanced Technologies for Digital Libraries, ECDL 2003 / Julien Masanès, Andreas Rauber, Gregory Cobena (eds). – URL: http://bibnum.bnf.fr/ecdl/2003/index.html

[10] Van Nuys, Carol (2003). The Paradigma project [online]. – *In:* RLG DigiNews. – Bd. 7, Nr. 2 – URL: http://www.rlg.org/preserv/diginews/v7_n2_feature2.html

31 July 2004