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-E
Meeting: 89. Cataloguing
Simultaneous Interpretation: -

The Paradigma Project and its Quest for Metadata Solutions and User Services

Carol van Nuys, Ketil Albertsen, Linda Pedersen and Asborg Stenstad

The Paradigma Project, The National Library of Norway

carol.vannuys@nb.no, ketil.albertsen@nb.no, linda.pedersen@nb.no, asborg.stenstad@nb.no

Abstract:

The National Library of Norway's Paradigma Project is working to ensure a satisfactory legal deposit of all types of digital documents – also the millions of documents found on the Norwegian Internet domain. Hopefully, Norway will be able to preserve its digital cultural heritage for the future, giving researchers access to an Internet archive by way of e.g. metadata and full text search. This paper gives a brief description of the project itself, before discussing the problems it encounters in its quest for metadata standards for discovery, long-term preservation, etc. The project's use of the FRBR entity levels work, expression, manifestation and item in the archive design will be presented, as well as ideas for future services: A verification and authentification service and an identifier allocation service - both available via the Internet.

1 Introduction

1.1 Web Archiving in Other Countries

Digital documents are disappearing daily, and one study[1] shows that only 20% of the documents found on the net, remain there – unchanged – after a year. Consequently, the possibilities for new generations of readers to study today's digital documents in the future are also disappearing. The preservation of our digital cultural heritage is an increasingly important and challenging issue, and the National Library of Norway[2] is just one of many institutions working systematically to find answers to the legal, technical and bibliographic problems that accompany it.

One facet of digital preservation work is to collect and archive documents from national Internet domains. Different countries have chosen different collection strategies: Denmark [1] and Australia [2] have taken the selective approach, while Sweden [3], Iceland and Finland have harvested their entire national web spaces. Norway is one of a handful of libraries in Europe that is harvesting and archiving digital documents from its national Internet domain based on existing legal deposit legislation[3].

1.2 The Legal Deposit Act

The purpose of the Legal Deposit Act [5] is to:

"[...] ensure that documents containing generally available information are deposited in national collections, so that these records of Norwegian cultural and social life may be preserved and made available as source material for purposes of research and documentation." (§ 1)

Considered extremely modern when it came in 1989, the present Legal Deposit Act, covers all *generally available* Norwegian documents stored on *any* medium: E.g., paper, microforms, photographs, combined documents, sound fixations, films, video, digital documents and broadcasting programs. Documents published abroad for Norwegian publishers and those specially adapted for a Norwegian public are also covered.

Of course, the World Wide Web had not yet appeared on the Internet in 1989. Digital documents – mostly in the form of databases – were few compared to today's millions of Internet publications, but they were still difficult to deal with technically. Today, the National Library's Long-Term Preservation Repository has the capacity to store 100 TBytes of data; the equivalent of a very large number of digital documents indeed.

2 The Paradigma Project

The National Library of Norway started the Paradigma Project[4] in August 2001. The project goal is to ensure the satisfactory legal deposit of Norwegian digital documents, and this includes the development of the technology, methodology and routines for the selection,

---

[1] Mannerheim, Johan. The WWW and our digital heritage [online]. - URL: Http://ifla.org/IV/ifla66/papers/158-157e.htm (Accessed April 15, 2004)

[2] For more information about the National Library of Norway, see URL: http://www.kb.nl/gabriel/libraries/pages_generated/no_en.html (Accessed April 15, 2004)

[3] Halgrímsson, Torsteinn (2003, februar 28). Web Archiving in Europe [discussion]. - NWA [online]. - E-mail-address: nwa@nb.no

[4] For more information about the Paradigma Project, see URL: http://www.nb.no/paradigma/eng_index.html (Accessed April 15, 2004)

collection, description and identification of *all* types digital documents – including those documents generally available on the Internet. The project is also working to give users access to its Internet archive in compliance with current legislation.

Project activities build on the Library's earlier work in several relevant areas, and four people are engaged fulltime. Approximately thirty staff members are also involved in project activities of some type. The project is scheduled to end December 31, 2004.

The following sections will briefly describe the project's ongoing work to select, collect and give access to digital legal deposit material from the Internet, as well as the nature and size of the Norwegian Internet domain.

2.1 Collection and Selection Strategies

2.1.1 Collection

Based on the Legal Deposit Act and recommendations from the Paradigma Project, the National Library has decided to start the general harvesting of *all generally available* digital documents from the Norwegian Web space (".no"). In time, documents found on domains such as ".com", ".org" and ".net", will also be harvested.

There are several reasons for taking this general harvesting approach: Firstly, we cannot predict which documents will be of value in future research and documentation, secondly, digital storage is becoming cheaper every day, thirdly, unfiltered harvesting saves resource-consuming manual selection at harvesting time, and finally, an Internet archive user can find documents via free text search facilities, thus being able to review all documents, including those that don't qualify for manual cataloguing. Selection criteria for any use, such as further bibliographic description, can also be challenged and changed at any time. This would, of course, be impossible if the material was excluded at harvesting time.

The Legal Deposit Section has harvested a selection of web documents semi-manually since 2001, using the HTTrack[5] software, and these documents are cataloged in the National Library's catalog (BIBSYS[6]). This activity will continue until the Paradigma Project's general harvesting activity and related procedures are fully established. The same section carries out event-based collecting as well, and it has collected, e.g., web sites belonging to political parties, prior to, during and after, elections. Other sections are also engaged in digital legal deposit activities, and the Library's Sound and Image Archive is working to find solutions for the legal deposit of "born digital" radio and television programs in cooperation with the Norwegian Broadcasting Corporation.

An extremely challenging issue is the deposit of the *deep web*, e.g., Internet newspapers, streaming media, documents from web cameras, interactive media and E-materials of all types stored in databases. The Paradigma Project has started the daily collection of approximately 65 Internet newspapers, and it will be downloading several entire newspaper databases in the near future, thus complementing the daily "snapshots". We are discussing deep web problems

[5] For more information about the HTTrack software, see URL: http://www.httrack.com/ (Accessed April 15, 2004)

[6] For more information about BIBSYS, see URL: http://www.bibsys.no/english.html (Accessed April 15, 2004)

within the framework of the *International Internet Preservation Consortium*[7], but a large number of administrative, legal and technical questions are as yet unsolved.

In summary, the National Library of Norway can expect to receive digital objects through several channels: Automated document harvesting from the Internet, database updates delivered in batches, subscription periodicals and mailing lists received through e-mail, NetNews discussion groups and documents delivered on physical media like CD-ROMs.

2.1.2 Selection

There are many valuable documents to be found on the Internet, and we are currently working to define *selection criteria* for those documents that we feel "deserve" manual bibliographic description at some level. These selection criteria are based on legal deposit legislation as well as the Library's general collection policy as formulated in our vision and strategic plan. Selection criteria for digital documents are being integrated with those for more traditional types in the Library's Selection Manual.

The Paradigma Project plans to implement a system architecture that allows a three-phase *selection* process, so that librarians receive technical help to find the few documents that should be cataloged at some level. The first phase finds and collects the Norwegian and Sami documents from the Internet. The second phase gives librarians the opportunity to *automatically* produce ranked lists based on specific queries. These lists are based on the use of vectors containing metadata that has been automatically extracted from the collected documents. In the third phase, librarians choose specific documents from the ranked lists for manual registration at some level, using the selection criteria mentioned above. Some day, we may also be able to monitor integrating resources that have been cataloged manually, thus helping librarians to discover and modify these bibliographic records e.g. at certain time intervals, when changes in the text exceed a certain per cent, etc.

2.2 The Norwegian Internet Domain

The exact size of the Norwegian Internet domain is still unknown at this time. The Paradigma Project's first harvesting round in December 2002/January 2003, resulted in approximately 3.1 million URLs (i.e. files), whereof approximately 53% (by count) are pictures (.jpg, .gif, .png). The NEDLIB-harvester[8] started with circa 1000 initial URLs, and harvesting was limited to the HTTP protocol, to the Norwegian national domain (".no"), and to URLs without parameters. The second harvesting round was carried out in August 2003, and it resulted in approximately 4.1 million URLs. The third harvesting round is currently underway, and no statistics are available at this time.

Assuming a distribution similar to that found in harvesting rounds conducted in Sweden and Finland, we expect to find 45-55% of the Norwegian Internet sites in domains outside of ".no. It goes without saying - manual handling and evaluation of each object is not possible; the vast majority must be processed automatically.

---

[7] For more information about this deep web activity, see URL: http://www.nla.gov.au/ntwkpubs/gw/66/html/p15a01.html (Accessed April 15, 2004)

[8] For more information about the NEDLIB harvester, see URL: http://www.csc.fi/sovellus/nedlib/ver11/documentation11.doc (Accessed April 15, 2004)

2.3 Access Strategy

2.3.1 *Who* will search for *What* in our Archive?

When trying to find metadata solutions for describing the rich and varied digital material stored in our archive, it is important to ask: Who will be using the material, and for what purpose? It is difficult to imagine researcher's specific questions in 10, 20 or 50 years, but we can try to imagine some *users groups* and *types* of questions.

One group may consist of users interested in studying the Internet and digital material as a *medium*, i.e. because the material has been gathered from the Internet and because it shows the characteristics of this medium. Here we can see that some users might need to study the use of language on the net and the relationship between different language forms; media researchers might want to study the relationship between printed and digital media or between technological development trends and content; users studying web page design may be interested in the use of advertisements, layout, etc.; researchers in the area of computer science may study different communication protocols, the use of formats over time and even data virus; social scientists may be interested in how the information available on the Internet has influenced society and visa versa. We can, of course, expect to find researchers with overlapping interest areas as well.

Another user group may consist of those needing to use digital documents as *source material* - just as they use traditional sources today. This group will most certainly consist of researchers from all subject areas, and it is therefore interesting to discover their expectations to digital material in particular. Is the relevant material available in digital form alone? Are dynamic content, animations, interactive displays, integrated sound- and video etc. of importance? Do researchers need to access material via free text search or correlate large amounts of information from different sources?

2.3.2 Current Legislation

Giving users access to the legal deposit Internet archive is a complex matter, and the National Library must find satisfactory solutions in spite of the many, and sometimes conflicting, regulations found in the Legal Deposit Act, the Copyright Act and the Personal Data Act.

We are currently trying to find answers to question like: Which users can receive access to different types of digital materials? Can they access the collections from computers outside the National Library?

2.3.3 Access Tools

User requirements like those described above are interesting to us, as we try to develop access tools for searching in our Internet archive. We must, of course, take into consideration the fact that librarians will catalog so few of the documents available there.

On a more technical plan, the Paradigma Project hopes to give users access to the Internet archive via the Nordic Web Archive's[9] (NWA) Access Tool (See Figure 1). Today, free text search with Boolean operators, search for a certain URL and presentation of document history via a timeline are standard options. Hopefully, the tool will also give us even more possibilities in the future: Use of Boolean search combinations to combine different hit lists, parallel search among cataloged documents in external bibliographic catalogs, searching in automatically extracted metadata, advanced programmed surfing and available pre-

[9] For more information about the Nordic Web Archive Project, see URL: http://nwa.nb.no/ (Accessed April 15, 2004)

programmed search parameters, options that let us store hit lists in a "project library", search access to document groups ranked according to different criteria (publisher, etc.), maximum one hit for existing duplicates, the grouping of a logical document consisting of many separate web pages as one hit, etc.

We plan to adapt the NWA Access Tool's interface to accommodate several special user functions, and our use of IFLA's *FRBR* model will play an important role in how we give access to archived material in the future.

3 A Quest for Metadata Solutions

The Paradigma Project is in the middle of its quest to find suitable metadata formats and solutions. The definition of metadata for *discovery* has been one of our main activities this past year, as well as our quest to find satisfactory solutions for the automatic extraction of technical metadata. In the following section we will attempt to give a little idea of *why* and *how* we plan to describe the many digital documents in our Internet archive.

3.1 Why should we catalog Internet Resources?

Nancy Olsen gives three basic reasons for *why* Internet resources should be catalogued in the introduction to her book *Cataloging Internet Resources* [3]:

1. There is a great deal of valuable information available through the Internet.
2. These resources need to be organized for accessibility.
3. Using existing library techniques and procedures and creating records for retrieval through existing online catalogs is the most efficient method of accessing these resources.

We agree with Olsen on all three points, but at the same time estimate that *far less than 1%* of the material collected from the Norwegian Internet domain may ever be subject to bibliographic registration at some level. This is, of course, because of the sheer magnitude of documents in the archive. We can try to comfort ourselves with the following thought: Although a much higher per cent of the Library's more traditional materials are subject to bibliographic registration, different materials are indeed treated in different ways: Ephemera is given a simple registration, while books and periodicals are given a higher level of cataloging.

In contrast, 100% of the Internet documents will be fully indexed with FAST[10] indexing software after harvesting. This will allow the Library's staff and users to search the Internet archive – both via free text as well as other indices. The tiny fraction of manually cataloged Internet documents will be available in full text from the archive and via bibliographic records in the Library's catalog – hopefully linked together in some user-friendly manner.

In addition to cataloging some documents, and indexing all documents, we will be harvesting existing embedded metadata as well as the Internet documents they describe, and the National Library is planning a future service that will allow publishers to generate and deliver metadata with their documents at the time of deposit.

---

[10] For more information FAST Search & Transfer (FAST) ASA, see URL: http://www.fast.no (Accessed April 15, 2004)

3.2 What is Metadata?

A quest for metadata solutions has, of course, led us to a quest for adequate definitions. The term "metadata" has been defined and redefined in the literature. "Data about data" is perhaps the most recurring definition, and metadata encompasses a whole range of information types[11]. We have discovered that metadata schemas are as abundant as they are diverse, but they do have one thing in common: They can help us *describe* and *find* the many valuable documents in our collection – also those that aren't candidates for high level cataloging.

3.3 What is an Internet Document?

3.3.1 Internet Document Definition from a Technical Viewpoint

When an Internet document is selected for harvesting and therefore archiving, the semantics of "a document" may be highly ambiguous: Which components should be harvested and archived as integral parts of the document? Which components should be subject to individual evaluation? We assume that any component affecting the "looks" (including *sound* and other *non-graphical* elements) of a web page unconditionally should be included if a web page is selected, i.e. background images, frame contents, images for buttons, etc.

Documents referenced through links are distinct from, but related to, the referencing document. At a higher semantic level, we often want to treat an entire group of documents linked together as one large document. If we treat them as fully independent documents, we run the risk of, say, harvesting a few chapters in a report, leaving other chapters out (this could be because they contain extended quotes, summaries etc., in other languages than Norwegian).

So, in answer to our question "what comprises an Internet document", we can say that an Internet document consists of many related parts or files, e.g. text, image, sound, animation, etc., and that these are most often connected by links and sometimes contained in frame sets.

3.3.2 Internet Document Definition from a Bibliographic Viewpoint

We can, of course, never rely on a computer to tell us where an Internet document starts and ends – even if we program it to follow certain instructions with this goal in mind. Luckily, librarians are very good at deciding which of the many parts of an Internet document comprise a logical whole. So, from a bibliographic viewpoint, we can define an Internet document as a unit of information that may be described bibliographically. This definition does *not* specify any set of definite or unique document components deliberately, but instead lets the librarian identify the object being described: An entire Web site may be described by one record, and one particular resource at that site may also be given a description. The librarian may include or omit background sounds, style sheets etc., and he may collect several closely related Web pages, e.g. chapters of one report, into one document. Our future automated procedures will suggest document definitions to the librarian, based on an analysis of the content, link classes, etc.: By default, embedded images, directly referenced sound/video clip and style sheets are included in the document. Links of certain classes, identifying a referenced Web page, e.g. as a table of contents or as a section, are also included.

---

[11] One of the many metadata surveys we have studied is: *DESIRE: A review of metadata: a survey of current resource description formats.* (1997). See URL:
http://www.ukoln.ac.uk/metadata/desire/overview/rev_toc.htm (Accessed April 15, 2004)

So, a unit of information that can be described bibliographically is the starting point for making a metadata description – both when digital material is deposited on fixed carriers like CD-ROMs, DVDs or when it is gathered as separate files from the Internet. This means that all digital documents – from *traditional* document types like monographs, dissertations, etc., *transient* document types like Internet newspapers, hyper poetry, hyper drama, etc. and *new* document types like homepages, web logs (i.e. blogs), etc. – are candidates for metadata description within the framework of our Internet archive.

3.4 A Metadata Survey and Related Work

3.4.1 Which Types of Metadata do we need?

We have found it interesting to ask which metadata formats the National Library uses today for the description of different types of digital materials. This information may be useful, as we someday hope to be able to import and export data to our archive. The results of our survey show that several formats are in use: BIBSYS-MARC (the BIBSYS system's MARC format) for digital text, Dublin Core Metadata Element Set[12] for radio programs, MAVIS[13] (an Australian system and format) for broadcasting material, sound and images, as well as other formats used in locally designed systems.

These metadata formats are well suited to their use, but they are not satisfactory solutions for all our metadata needs. The Internet archive needs many types of metadata: *Administrative* metadata regarding e.g. the creation and modification of metadata records, *rights and access management* metadata to store copyright information and define which user groups can gain access to the archive and which documents they can read, s*tructural* metadata for showing logical relationships between objects, between metadata or between objects and metadata, *long-term preservation* metadata for the specification of e.g. file types, necessary software and document conversion/migration history, and finally, *technical* metadata for specifying the documents size, scripts, communication details, etc. Last, but not least, we need *descriptive* and *analytical* metadata for search and retrieval purposes.

3.4.2 Which Description Model should we Choose?

There are several opinions as to which level of description a digital document should receive. In our work to define metadata for descriptive and analytical metadata, we have looked at two alternative models. One alternative is to use three description levels:

1. Cataloging for inclusion in the National Bibliography/the National Library's catalog BIBSYS/other special databases.
2. Cataloging at a simpler level in a common format.
3. An automatic extraction of metadata from the document itself as well as from communication protocols, etc.

The other alternative is to use a two-leveled model, i.e. "to catalog – or not to catalog":

1. Cataloging for inclusion in the National Bibliography/the National Library's catalog BIBSYS/other special databases

---

[12] For more information about Dublin Core Metadata Initiative, see URL: http://www.dublincore.org (Accessed April 15, 2004)

[13] For more information about Wizard's MAVIS system, see URL: http://www.wizardis.com.au/ie4/products/mavis/introducingmavis.html (Accessed April 15, 2004)

2. An automatic extraction of metadata from the document itself as well as from communication protocols, etc.

There are several arguments for this second alternative: 1) Retrieval of digital material (free text, etc.) is not dependent on registration as is the case with unregistered analog material. 2) It is unnecessary for the Library to register material in order to keep track of its logistics, e.g. which university libraries has received copies. 3) We can always regret our decision not to catalog a certain type of digital material.

A brief description of each of the three levels is given in the following section.

- **Cataloging for inclusion in the National Bibliography, etc.**

At this time, our suggestions for which document types should be cataloged at this highest level are incomplete, but we can say with certainty that a small number of valuable digital documents will continue to be cataloged in some MARC format for inclusion in the National Bibliography. (We can mention that Norway's version of MARC is called NORMARC, that some systems have adopted local versions, e.g. BIBSYS MARC, and that the use of MARC21[14] is being discussed at a national level. Norway's cataloging code is based on the second edition of Anglo-American Cataloguing Rules (AACR2), and Chapters 9 and 12 are now available in Norwegian.)

We can also say with certainty, that the cataloging of audio-visual materials for long-term preservation requires a high level of detail – especially when it comes to keeping track of technical information connected to the restoration of originals, copies, etc. The Library will undoubtedly continue to use MAVIS for this work.

- **Cataloging at a simpler level in a common format**

As earlier mentioned, the National Library is planning a future service that will allow publishers to generate and deliver metadata with their documents at the time of deposit. Today, the Paradigma Project is working to define the metadata format(s) that will form the foundation of a future user-friendly tool provided by this service. Eventually, librarians may handle the metadata records supplied by publishers, using these as the basis for higher-level bibliographic records.

We have analyzed and compared a few metadata formats in our work to find suitable solutions: MAchine Readable Cataloguing (MARC) and Dublin Core Metadata Element Set (DCMES), as these both are used in libraries and related institutions; Metadata Object Description Schema (MODS)[15] and Metadata Encoding & Transmission Standard (METS)[16], as these have been developed by libraries for the library community and Online Information eXchange (ONIX)[17], as this format is developed by the publishing and book industries. We also note that the ISBN community has suggested that in the future, registrants may supply ISBN agencies with ONIX compatible metadata in connection with the assignment of each ISBN.

We have compared the formats above by asking: Who is responsible for managing the format? Is it an international standard? In which area is it used? What type of media does it

---

[14] For more information about MARC21, see URL: http://www.loc.gov/marc/bibliographic/ecbdhome.html (Accessed April 15, 2004)

[15] For more information about MODS, see URL:http://www.loc.gov/standards/mods/ (Accessed April 15, 2004)

[16] For more information about METS, see URL: http://www.loc.gov/standards/mets/ (Accessed April 15, 2004)

[17] For more information about ONIX, see URL: http://www.loc.gov/standards/mets/ (Accessed April 15, 2004)

describe? Does it include semantic and/or syntactic definitions? How does it describe the relationships that exist between documents? Is the format dependent on specific rules or codes? Is it compatible or related to other formats? How widely is it used, and by which communities?

We hope this survey can lead to a broader metadata discussion in the Library in connection with its ongoing review. We also plan to look more closely at how we can satisfy our user's functional requirements by using *Common core records* proposed by the IFLA Working Group on the Use of Metadata Schema [6] and IFLA's *Final Report on Functional Requirements for Bibliographic Records* (FRBR) [5]. Collaboration on metadata solutions with an ongoing bibliographic project within the National Library itself, as well as with *The Norwegian Digital Library*, another project at the national level, is also on our agenda. Finally, we hope that this work will result in recommended metadata formats for description at different levels.

In the mean time, we have worked to specify technical metadata requirements for our archive system software. We have identified several factors that can influence our choice of metadata format for lower level cataloging. Here are a few technically desirable factors:

- Semantic-interoperability with MARC: It is important that the metadata format's attributes are semantically harmonized with the library community's dominating MARC format. If possible, the format should be a functional subset of MARC. This would facilitate the exchange of data.
- Simple yet rich: It is important to find a metadata format that is simple to use, yet rich enough to let us represent an adequate amount of detail.
- Easy to convert to other formats: Conversion crosswalks between the chosen format and MARC should be available or relatively easy to define. Here we see that crosswalks between MARC21 and MODS - and between MARC21 and ONIX - already exist, as well as a crosswalk between unqualified Dublin Core and MODS.
- XML-compatibility: XML is more or less a de facto standard, and a format that is XML-compatible, will let us handle the format with available software. A larger framework structure will also be defined in XML, thus letting the archive accept metadata from different sources, handle metadata modification, define original metadata, keep track of version history, etc. (e.g. METS).
- Extensibility: A metadata format should let us define new elements when necessary.
- Core elements: It is important to define core metadata elements, i.e. a common denominator that can facilitate document search and retrieval between different material types.

If we compare these factors with the metadata formats described in our survey, we see that formats that are MARC and XML compatible are preferable. However, there is no simple recipe. New elements for technical, structural and rights and access management metadata must be defined, and perhaps consolidated within the METS framework. The same, of course, is true in regard to metadata for long-term preservation. Here the Library's Long-Term Digital Repository requires us to use OAIS[18] compliant metadata.

---

[18] For more information about OAIS Reference Model, see URL:http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf (Accessed April 15, 2004)

- **An automatic extraction of metadata**

Unfortunately, librarians will never catalog a mind boggling 99% of the Internet documents in our archive. This is why we are currently investigating the use of automatic analysis and extraction of metadata from Internet documents as part of our work with metadata and system design. Extracted metadata will be stored together with the digital objects and other metadata descriptions, and it will be made available for structured searching in the Internet archive.

The technology is not yet good enough to decide a document's type automatically, but it can help to reduce the amount of documents that are given human attention in phase two of our selection work. Examples of such document type properties are 1) language, vocabulary and grammar, 2) document size and structure, 3) source/publisher/web-server, 4) use of ”cookies” 5) given age and life expectancy of a document, 6) sound, pictures, animations, video and other advanced types of information, 7) user interaction like “forms”, buttons, etc., 8) number, type and source of links, 9) URL-values, e.g. use of special words or characters in the URL, 10) use of client-side scripts, 11) technical communication details.

The technology for analyzing vocabulary and grammar is improving, and we feel that this type of analysis may be an important element in future automatic procedures. Eventually, automatically chosen type properties will be made available for structured searching in the Internet archive. The value of these properties will be limited, but in combination with other search criteria, they may indeed prove to be useful.

4 FRBR's Role in the Internet Archive

The Paradigma Project wants to present the archived digital documents and metadata in an organized and structured way, thus facilitating user navigation. We have found IFLA's FRBR model to be an essential tool in this work, and we will be using the model as a foundation for the design of the Internet archive.

We believe that adding aggregate modeling mechanisms to the FRBR model will benefit our work with dynamic media such as Internet documents, multimedia and other continuing resources. Aggregate mechanisms can be implemented as pure extensions to the model, requiring no significant changes to the existing FRBR concepts. An article on our proposed aggregate mechanisms will be made available some time this year in the FRBR theme issue of *Cataloging & Classification Quarterly*.

To adapt the FRBR model to dynamic Internet documents, a moderate reinterpretation of *manifestation* and *item* level concepts is required, and these are described in the following section.

4.1 Adaptations of FRBR for Use with Dynamic Internet Documents

4.1.1 Dynamic Documents

Internet documents are often *dynamic*, e.g. an Internet newspaper, updated many times a day. A user may relate to this type of dynamic document as a forum or information channel: “*The Daily News* reports that...” We can perhaps say that a dynamic document corresponds roughly to an URL. Concepts of ”issues” and successive ”editions” must aslo be re-thought in an Internet context: From a formal viewpoint, a Web page update may be similar to a new book

edition. Yet, readers view e.g. the continually changing front page of a Internet newspaper as a single, changing entity - not as distinct, separate editions.

Using the FRBR model with extensions for aggregate components, we have defined the concept *dynamic document* as "the entire life cycle of a continuously changing Web page or similar Internet document".

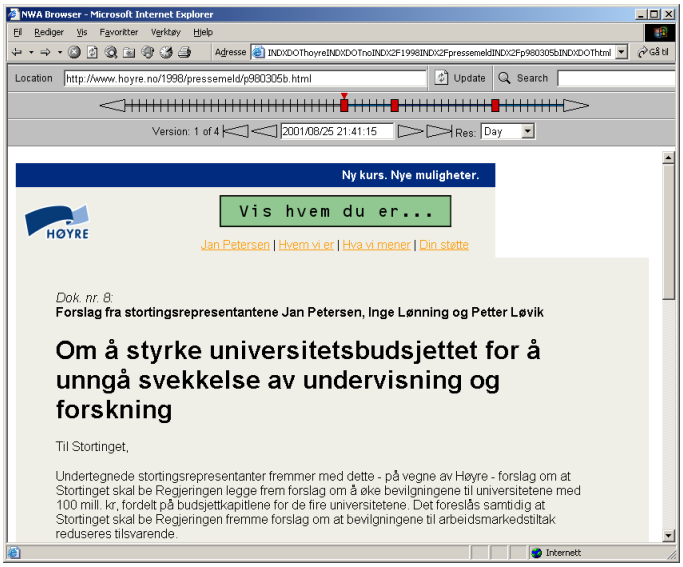
If we were to catalog an updating Web document of this type according to AACR2, we would normally use rules for integrating resources, i.e. a bibliographic resource that is added to or changed by means of updates that do not remain discrete and are integrated into the whole. But, documents like Internet newspapers are more like a radio channel, a constantly changing flow of transient information. They do not "integrate into the whole". Capturing the contents of a continuously changing document at a given moment is like recording a *sample* of an ephemeral broadcast. We term each such sample or snapshot a *specific document.*

When a dynamic document is accessed on the Web, the *item* (i.e. exemplification) retrieved by a user, may be different from all other *items* of the same document: It may depend on a combination of a number of factors: User identity, the access tool used (Web browser), information about earlier accesses to the same document (preserved in cookies), parameters explicitly specified by the user e.g. in a form, and, last but not least, the current state of a database. Often, the *item* is generated on the fly when a user requests an exemplification. In other words, an HTTP call acts like a "print on demand"-service: The copy delivered reflects whatever the content of the document database is at the time of printing. The database may be considered to be a (semi-) permanent physical representation of the dynamic document, from which specific *items* may be derived. The *items* themselves have no permanent representation - they are transient unless preserved e.g. in an Internet archive.

4.1.2 Specific Documents

We have defined an *item* exemplifying a dynamic document as a *specific document,* differing from a traditional *item* in one primary respect: It is a member of a group of *items* exemplifying the same dynamic document. A document stored in the archive, or displayed to the user, is obviously a specific document, but this is toned down: A full text search will give at most *one* entry in the hit list for a dynamic document. If the user requests a display of a hit, the dynamic document is presented as one unit, and the user can then select a specific *item* on a *timeline*, i.e. a menu line representing the document's lifespan. Each preserved version, i.e. specific document, is indicated on this timeline with a marker. The user may access any specific document by clicking the marker for a certain date/time, thus retrieving the *item* (See Figure 1).

FIGURE 1. Presentation of a dynamic document in the NWA Access Tool User Interface



4.2 Publisher or User Defined Document and Metadata Definitions

The presentation of archived documents for purposes of research and documentation is just one service that will be provided by the National Library. In addition, and based on the ideas presented above, we have suggested revisions to the Library's existing identifier allocation service. Today, this Web service assigns URN:NBNs [7] to universities and other institutions from the Norwegian branch of the URN:NBN name space. We can, however, see possibilities for allocating stand-alone ISBNs from this service as well.

4.2.1 Future Functionality – a Scenario

A scenario showing future functionality is as follows: The primary identifier series assigned by this service requires the user/requestor to supply both a minimum set of *metadata* and an exact definition of the identified document.

Work, *expression*, *manifestation* (including dynamic document definitions) and *item* (including specific document definitions) identifiers may be allocated. *Items* (specific documents) must be specified by a complete component list (e.g. an HTML file, picture files, sound files, etc.); *manifestations* (dynamic documents) may also be specified by rules, such as "The Internet newspaper front page at this URL and all pages directly linked from the front page that resides on the same Web site".

For *expression* and *work* identifiers, the user may optionally identify *expressions*/dynamic documents and *items*/specific documents, which are instantiations of this *work*/*expression*.

Publisher or user defined definitions are considered final rather than automatically proposed. The identity of the publisher or user allocating the identifier is archived; a document definition specified by a recognized publishing house or university may be considered more significant than one requested by an arbitrary user.

4.2.2 Metadata Fields

Obligatory and non-obligatory metadata fields could be available for the description of the document at each FRBR level, and each level would be identified with an URN:NBN. The

metadata values will be stored with the identifier, and users of our Internet-based resolution service will be able to find the document based on this number.

After filling information in the metadata fields of a future metadata/identifier allocation tool, it would be possible for a publisher to click on a button, e.g. <HTML Dublin Core>, in order to view this metadata in HTML in a separate window. The user could then copy the metadata and paste it in the <HEAD> element of the Web document being described, before continuing the identifier allocation process. After saving the digital document, now including embedded metadata, the user could easily store the metadata enriched document copy in the Library's archive by clicking the browser's update button.

4.3 Possible Document Verification and Authentification Services?

We have heard tales of authorities that revise official statements on the Internet, and then later refuse to acknowledge the existence of earlier versions. We have also heard of commercial firms that advertise their products at a certain price, and then bill the customer for a much larger sum.

With these and other stories in mind, the Paradigma Project proposes a verification and authentification service that could let users request a download of a given Internet document, i.e. a snapshot of a web page containing a particular commercial offering, statement of legal responsibility, libel, etc.

If doubts should later arise with respect to these documents' content at a given time, the Library could then confirm (or reject) any claims in this regard. Even when no legal aspects are involved, a preserved specific document *item* may serve as a well-defined image of a dynamic document at a given time, e.g. for quoting or referencing purposes. This is important, especially when we realize that most Internet documents have no page numbers, no version number, etc.

In our Internet archive, a specific document is defined in the form in which it was received from the Web server. There is a well-defined bit stream for each component of the document (text, pictures, etc.). The graphic rendering of the document is *not* part of its definition – this process is left to the access tool. The specific document is identified as the content of a dynamic document given by certain components and metadata:

- the source of each component (e.g. a URL)
- all parameters specified by the client when retrieving the components
- the wall clock time when each component was retrieved
- the set of components included in the document

5 Conclusion

The National Library of Norway's Paradigma Project is working hard to establish satisfactory technology, methodology and rutines for the legal deposit of all types of digital documents – also the millions of documents found on the Norwegian Internet domain – within the remaining project period. We hope to be able to give our users access to archived material via bibliographic records, diverse types of metadata and full text searching tools already in 2005.

Our FRBR structured Internet archive will most certainly be one of the first of its kind, and we also hope to realize our ideas for the use of the FRBR entity levels *work*, *expression*, *manifestation* and *item* in a future identifier allocation service. Perhaps our ideas for a

verification/authentification service on the Internet will also become reality some time in the future? Time will tell, but in the mean while, the National Library will continue to explore new ways to preserve Norway's digital cultural heritage and to provide its users with tools that can open the doors to this exciting digital library.

Selected References

(All URLs were accessed on April 15, 2004.)

[1] Final Report for the Pilot project "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] Norway. [The Legal Deposit Act (1989)] Act relating to the legal deposit of generally available documents : no. 32 of 9 June 1989 : with regulations / [published by the Ministry of Church and Cultural Affairs ; unofficial English translation published by the National Library of Norway. - [Oslo] : National Library of Norway, 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. – *I*: 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. – *I*: 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]. – *I*: RLG DigiNews. - Vol. 7, no. 2. – URL: http://www.rlg.org/preserv/diginews/v7_n2_feature2.html

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