Identification of Network Accessible Documents: Problem Areas and Suggested Solutions

Carol van Nuys, Ketil Albertsen

The *Paradigma* Project at the National Library of Norway

Abstract. Network accessible documents may be highly dynamic entities; developing over time and frequently changing location. This paper discusses issues raised by dynamic behavior and properties particular to such documents, and proposes methods for handling problems that are considered important, with particular emphasis on legal deposit documents. It also presents a proposal for a scheme for the unique identification of network accessible documents, for retrieval and citational purposes, and for a revised Identifier Allocation Service design.

1 Introduction

The Internet as an information source. For years, the Internet has served the scientific community as a primary source of up-to-date information, and as a distribution channel for research results. This trend is rapidly spreading to all academic fields, and Internet use is no longer restricted to universities. Today the net is used as an integral tool in learning situations at all levels.

The Internet is rapidly being established as the *first* source of information: Users carry out initial problem surveys and read about the latest developments. Books, and even journals, more suitable for background material or in-depth studies, are often consulted after initial Internet surveys are completed.

Unfortunately, dynamic documents found on the Internet disappear daily. This is also the case for cited network accessible documents (hereafter network documents).

The Paradigma Project.[1] The Norwegian Legal Deposit Law[2], requests the deposit of network documents – also those located on the Norwegian Internet domain. The National Library of Norway initiated the Paradigma Project in August 2001. The project's main objective is to establish procedures for selection, registration and long-term archival of, and access to, all types of digital documents. Many problems often overlooked in interactive browsing situations, appear within the framework of legal deposit: E.g., the need for downloading and identifying streaming information, rights and privacy issues, and functions for document authentication and version handling.

The need for unique identification. An unambiguous identification of network documents is required for several reasons: An author must be able to quote and reference work from previously published documents. A reader, upon encountering such a reference, must be able to retrieve this cited document by using the given

reference. A person or institution offering information to the public needs a unique reference for use in marketing or public relations work, as well as unambiguous identification, if the information has contractual or other legal aspects.

Current state. Today the only widespread identification (ID) scheme for referencing network documents is the World Wide Web *URL*[3] (Uniform Resource Locator). URLs were never intended to identify document *content*, but rather – as the name implies – its *location*. Superficially, URLs may appear suitable even for content identification, but a number of problem areas exist. Some of these will be subject to further exploration in this paper.

Several alternative content ID schemes are proposed and, to varying degrees, implemented; DOI[4], PURL[5] and URN[6] are noteworthy examples. The schemes commonly establish procedures for assignment and management of ID values, often with procedures for associating metadata with the identifier. In most cases, important *semantic* aspects of an ID are left unspecified. A registrant (i.e. a person/an organization defining an identifier/object relationship) is often free to imply any semantic meaning beyond what the ID scheme explicitly addresses.

2 Problem Areas

While some schemes do address one or more of the issues listed below, we know of no existing scheme that properly handles *all*, or even a significant fraction of these issues with respect to network documents.

2.1 Specificity of Reference and Abstraction Levels

Authors frequently have a limited awareness of abstraction levels. Too general or too specific references are both commonplace. In manual handling of documents, few problems arise: A bookstore clerk is able to find the paperback edition for a customer even if he only knows the ISBN of the hardbound edition – a too specific reference. The newsstand man correctly interprets a request for "Today's News", even though the reference is highly ambiguous, considering the 300 or more different newspaper issues published each year.

Automatic, computer-based systems on the other hand, lack some or all facilities for requesting additional data from the user. The search key presented to the system should, implicitly or explicitly, reflect something about the desired specificity.

IFLA's FRBR Model. The dynamic nature of most network documents' content, availability, location and form, makes traditional referencing methods, e.g., by ISBN, less than suitable. IFLA's FRBR Model[7] (FRBR), addresses a number of identification related issues, and our thoughts are based on this model.

FRBR defines four abstraction levels for the description of products of intellectual or artistic endeavor: *Work*, *Expression*, *Manifestation* and *Item*. Documents may be referenced at these varying levels of abstraction, ranging from the abstract ideas embedded in a work, to one specific, physical copy of a specific edition.

FRBR Example. The Norwegian Board of Health publishes instructions for use of antibiotics in hospitals. The *Work* level description of this document may be supplied with a uniform title, and it says nothing about the document's language, edition or binding. This may be the most appropriate identification level in an international environment.

When choosing an edition of Board of Health instructions, the *Expression* level description may be used. The user may perhaps select the original Norwegian document titled *'Smittevernloven : håndbok : bruk av antibiotika i sykehus'*.

A *Manifestation* of the Norwegian title is available in MS-Word, RTF, PDF and HTML formats, and a reader selects *one* of these for downloading or viewing. This *manifestation* level is identified with the series number *IK-2737*. A researcher making references to this report should be unconcerned about the format selected by the reader; he should *not* refer to, say, the RTF version.

In the manifestation's preface, hospitals are encouraged to make local amendments and adaptations of the report to suit local conditions. Hospital personnel will, in the course of their daily work, need a reference to this *Item*, i.e. the specific modified copy of the document.

Only *Manifestation* and *Item* are given specific identifiers in FRBR, but giving identifiers to each of these four levels can be fully justified. By choosing a given level, a referencing author conveys his intent and says what he considers to be significant: The generally abstracted idea alone, the way the idea was expressed, the presentation of the idea in a given format, or one specific unit.

FRBR is not yet fully developed for network documents. The *Work* and *Expression* concepts apply without significant adaptations in the Internet world, but the semantics of *Manifestation* and *Item* need reconsideration. In our view, a useful interpretation of the two lower FRBR levels is to view a *Dynamic document*, such as a web newspaper under constant revision, yet addressed through one URL, as a *Manifestation*. A snapshot of the newspaper, defined here as a *Specific document*, is analogous to the *Item.* We agree that there is a definite need for identifiers at both these levels: The distinction between bit-wise identical, digital copies of a network document has little or no semantic significance, and their identification is of secondary importance with respect to persistent, globally known identifiers.

Introducing distinct identifiers for the different FRBR abstraction levels raises two sets of problems: The task of creating a complete definition of format and semantics, and the task of teaching referencing authors – and other user groups – how to select and use the proper kind of reference. Both sets of problem are as yet unsolved.

2.2 Completeness of Information

The information available in traditional documents is obligatory: When you get the document, you get it all – binding, choice of typeface and page layout, table of contents, figures etc. Network documents, however, are different: When viewing or downloading a document, a user can choose to ignore scripts, background sounds, animations etc. and instead copy & paste textual content. In other situations, a user might be interested in graphics or sound alone, and ignore everything else.

The complete information available in digital documents can also be made obligatory for download/viewing. This rather naive approach can lead to several problems, e.g., layout parameters can make the information less accessible for some users, the data volume transmitted to the user can be unnecessarily large or a higher degree of consent from right holders might be necessary.

Users want, need or may be permitted access to, varying degrees of "information completeness" for a digital document. A given reference should be semantically unambiguous; different levels of completeness should be expressed as different ID values. By selecting a certain value, a referencing author signals which aspects of the referenced document are semantically significant in the referencing context.

2.3 Structured Documents: Extent and Granularity of Identifier

Identifier semantics should indicate the document's extent: When *downloading* a network document, the ID should cover the *entire set* of document objects. When *displaying* an HTML[8] document, the ID should cover objects included in the first web page only – other identifiers should cover subsequent pages and objects.

For composite objects, an ID may cover the *references* to components, or treat the *values* of the components as an integral part of the identified composite object.

In a research situation, when referencing/quoting documents, we may need to identify smaller units than an entire document: A specific figure, table, chapter, page etc. Traditionally, this is done by informal textual specification. In a digital world however, a reference that can be interpreted automatically, provides great advantages: The reader can navigate directly to the referenced point in the document, the referencing author can, given appropriate tools, insert exact references semi-automatically, e.g., by "drag & drop" techniques, the references can be automatically verified and a right holder can release a reference to selected parts of a document.

At its lowest level of granularity, a reference may identify a *point* in the document, or an *extent* (starting point plus length or ending point). We term these *point references* and *fragment references*, respectively.

HTML provides "logical" point references, but these require the referencing author to insert *anchors* at certain positions for use in external referencing. This is clearly unsatisfactory as a general solution. *SICI*[9] (Serial Item and Contribution Identifier) suggests another solution. Here the referencing author may construct an ID value. The complexity of the SICI specification reflects the complexity of the problem area. Still, although suited for identifying journal contributions, SICI is far from complete for general use, e.g., for network documents. *PII*[10] (Publisher Item Identifier) has a simpler syntax, but its application scope is not significantly broader.

In order to be fully unambiguous across different formats and abstraction levels, point/fragment references must share a common, externally visible granularity (i.e. structure, decomposition). Otherwise, the precision of the reference may be less than perfect in some representations.

A fragment reference mechanism may include facilities for accessing rights: A right holder may permit general access to a fragment of a document, but not the whole. E.g., a record company may allow any user free network access to the first 30

seconds of each track on the company's CDs. A research institute may grant more liberal access to the summary chapter of a report than to the main text body.

2.4 Identifier Format

A document ID, displayed e.g., in a list of references, may be "structured" along the lines of SICI, conveying some meaningful information about the document or point/fragment. Or, the identifier may be "opaque", thus conveying no meaningful information. This identifier must be "resolved" to a document or fragment by a computerized system.

A resolution mechanism may map opaque identifiers dynamically, depending on external conditions, such as the reader's preferred format. If the referencing author has prepared mappings for different abstraction levels and formats, the reader may see a logical reference independent of physical/logical format (within the limits permitted by each format). Opaque values are generally known to be more compact than structured values; this simplifies keyboard entry and may improve readability.

Structured values depend less on resolution facilities for interpretation, and may be more "user friendly". It is easier to build tools for finding/building correct references; the software can e.g., be aware of existing and possible variants. Values are often taken from a sparsely populated domain; typing errors are usually detected without the use of check digits when the ID is resolved.

2.5 Identifier Persistency

An identifier is persistent when it identifies the same object for an indefinite period of time. The resolution services for most identification schemes may still be operative 20 years from today – the "far" future is not a pressing concern. A deposit library, however, will need to map IDs to documents 100 years from today, and having to rely on external mapping services is not a satisfactory solution. It is necessary to take the proper steps to ensure that IDs can be resolved in the distant future.

A well-designed, structured scheme is less dependent on complex resolution than an opaque one is. However, there is a general trend towards the use of opaque schemes (e.g., DOI); this trend includes changing previously structured schemes to opaque ones (e.g., new schemes for Norwegian phone- and car license plate numbers, proposed, although rejected changes to ISBN).

Identifier persistence requires that an identifier can be resolved even after the object itself no longer exists. This means that a user may be able to find a document replacement, such as a newer version or a variant in a different format.

2.6 Document Content Persistency

Identifier persistency says nothing about the persistency of the document's *content.* Documents, both in the physical and network worlds, can go through major changes while still retaining their identity. Most schemes do not clearly and unambiguously

define which changes in content are permitted while still retaining the object's ID. Exact definitions of content change may be essential for schemes that are used to identify legal documents. This may also be the case when identifying documents in an archival or library context, e.g., whether or not to archive a new version of the object.

2.7 Identifying Chronological Document Versions

If a structured identifier scheme assumes that content revision beyond a certain threshold requires a new ID, the two ID values may overlap or share parts. Changes could be reflected in part of the identifier, e.g., a "version number" or "date".

An opaque ID cannot convey version semantics; IDs for versions of the same document are independent of each other. Relationships between versions must be represented by other mechanisms.

2.8 Identifying Parallel Variants of a Document

In a structured scheme, alternate content, expressed as, say, different language editions (*Expression*) or different file formats (*Manifestation*), may be assigned partially identical ID values, differing only in e.g., a "format indication" field. Some implementations however hide this for the user; e.g., a web site may use the submitted browser ID as an URL extension to select a variant tailored to the browser.

An opaque ID cannot convey variant semantics; variant identifiers are independent of each other. Association between variants must be represented by other mechanisms.

2.9 Uniqueness of Document Identifier

A scheme may assume that a document is assigned one and only one ID; "synonyms" are not allowed. Other schemes allow limited or arbitrary synonyms; e.g., a book may under certain conditions be assigned two or more ISBNs.

Partial document overlap is common on the web – graphics, sound/video clips and style sheets are shared among documents. Hierarchical schemes defined according to document structure must allow multiple identifiers for common components.

Multiple identifiers may improve recall. If the externally visible identifier is used to determine the location of the document in the archive, a single ID must be primary. Other identifiers – if permitted – must be treated as secondary keys. Alternately, the archival system may treat *all* externally visible identifiers as secondary keys, assigning an internal primary key for the archive location.

2.10 Identical Content and Conflicting Updates

Identical web documents are often found at multiple locations (URLs); an ID intended to identify the content should be the same for all copies. Updates (new versions) of the copies may be unsynchronized – one copy is updated first, followed by update

propagation to other copies. The status of the copies is ambiguous: A not-yet-updated copy may be seen as an authoritative copy of the previous version, as invalidated, or as a non-authoritative but valid copy of the previous version.

A site may skip updates, creating holes in the version sequence when compared to other sites, but invisible when the site is viewed in isolation. The version sequence may be considered local to each copy, hence continuous even if versions are skipped.

The site first updated may change from one version to the next, making it difficult to identify one definite authoritative document copy.

Copies may start out as identical, and remain so for several versions, after which *different* updates are made. There are no definite rules for deciding which copy retains the original ID, or how to assign new IDs to other copies.

When using opaque primary IDs, these problems are not visible in the identifier value, but they do appear in any identifying secondary key scheme or other version-aware structure.

2.11 Human Aspects in Manual Handling of Identifier Values

Human readability. An identifier should have a format that makes it easy for humans to handle: It should either be short, or its value should be meaningful to the reader. Symbolic names are usually considered more meaningful than numeric codes, which are more meaningful than arbitrary character sequences (such as those produced by MD5[11]). The readability of long numeric identifiers can be significantly improved by splitting the digit sequence into parts using separators, such as the use of hyphens in ISBN and ISSN numbers, even when the separators have no semantic effect and the parts/fields created may represent nothing at all.

Error detection. An identifier, e.g., in a list of references, must be suited for manual keyboard entry. The scheme should aid in detecting typing and reading errors, e.g., by defining a sparsely populated domain, or through check digits.

For a dense, opaque scheme, minimum-level error detection can be to define a standard ID length, so missing or superfluous digits/characters can be detected at entry time. Check digits provide higher quality entry time error detection.

Structured schemes are usually sparsely populated. Without check digits, typing errors cannot be detected at entry time; a resolution operation must be used to reveal errors. This may be considered sufficient – if not, even a structured ID can be augmented with check digits.

2.12 Document Authentication

Modifying a digital document is far easier than modifying a paper copy. It may be difficult to determine which one, of two or more different copies, is the “original”. This may have major legal consequences if the document in question is, say, a business contract, bill of sale etc. Digital signatures provide a partial solution, but are not applicable for large application areas. Examples are product prices claimed in net ads, or libel in a publicly available web page; these are rarely digitally signed.

An authentication service may be provided as a supplement to a resolution service, or may be handled by specialized, authorized institutions. The service must be capable of handling non-public documents as well: It cannot unconditionally make the authoritative document available to the requester, but it should accept a copy of a document *claimed* to be identical to the authoritative copy. The service should return the result of a comparison, either as a simple true/false value, or indicate roughly the amount of difference. Specific differences should *not* be returned, as this might reveal confidential document content.

3 A Design for Network Document Identifier Assignment

The National Library of Norway has for some time been running a pilot web service, offering publishers unique identifiers for printed and digital documents in the Norwegian branch of the URN:NBN[11] (Universal Resource Name : National Bibliographic Number) name space. The primary extensions suggested in this chapter by the Project include a closer specification of the semantics of a given ID. The ID syntax is extended with check digits; the use of check digits may be limited to presentation in a user interface and in a printed (offline) representation of the ID. Decisions about these suggestions are not yet handled within the National Library of Norway.

3.1 Registrants and ID Availability

Paradigma suggests that the ID Allocation Service be made generally available: E.g., any publisher, author, researcher or librarian needing to reference a digital source, should be able to use this service to request an ID for a document, a point in, or fragment of, a document, whether they have intellectual rights to the document's content or not. Any reader should also be able to obtain an unambiguous reference, e.g., for a citation or legal purposes, without the author's consent or involvement.

An implication of this is that for any given network document, a number of IDs may be registered. These usually represent different document versions or variants. This is by design; the IDs all represent published, well-defined document versions.

Certain services should be available only to recognized registrants such as publishing houses. Recognized registrants must identify themselves to the service with a name and password.

If an ID is requested for a document that already has been assigned one, the registrant is made aware of this and may cancel the request, and instead use the existing ID. If the request was made before the document was evaluated for cataloguing at the National Library, the document extent, information completeness, metadata etc. specified by the registrant may be presented to the librarian if and when the document is considered for cataloguing. Information from a recognized registrant, such as a publisher, takes precedence over information from non-recognized registrants. If the document is catalogued, the ID may be included in the bibliographic information.

An index of allocated document and point/fragment ID values can be made generally available online. Availability of the ID and its definition does *not* imply that

the document *content* can be retrieved from the National Library; legal deposit documents are not available to the general public, even if they were earlier generally available on the Internet.

If authentication is provided for Internet documents, this service could also be offered for documents other than legal deposit ones: Any document, possibly containing confidential information, could be submitted for identification and registration. Whether this service should be offered for non-published documents is a policy question that has not yet been discussed. A decision will not be made until the legal status of the service has been clarified.

The allocation of an ID may give reason to create an internal digital object in the National Library's archive, an object that represents the identified entity: A *container* object collects components e.g., of a defined document, an *agent* object represents an abstract entity, a *reference* object is a logical reference. These objects are the holders of the assigned ID values.

Metadata may be registered for abstract entities represented by an agent object (as well as for concrete network documents). We distinguish here between an agent object that exists at the object level, and a bibliographical object that exists at the metadata level.

3.2 Identifier Format, Component and Document Relationships

In limited contexts, structured IDs have desirable readability properties. However, a scheme that is sufficiently general to cover all media types and object classes will be extremely complex, and it may lead to very long, hence less readable ID values.

Opaque values allow far more flexible adaptation to various object structures, including sound, pictures and film/video, where relationships between objects and their components form non-hierarchical relationships. ID values are usually short and therefore readable, and better suited for addition of check digits.

The Paradigma project therefore suggests an opaque ID, consisting of a fixed-value alphabetic string prefix, a densely allocated numeric value, and two trailing check digits, which facilitate entry time error checking.

Versions and variants of documents can be defined as groups using container objects. Elements in containers may be ordered, for versions, or unordered, for parallel variants.

3.2 Abstraction Level

The registrant indicates the abstraction level, i.e., the specificity of the document reference, as one of the four levels defined by IFLA's FRBR Model: Work, Expression, Dynamic Document (Manifestation) or Specific Document (Item). Appropriate interpretation of the two lower levels for network documents, as discussed above, applies here.

We suggest that the registrant must specify a minimum set of metadata before receiving an ID. If the allocation service is already aware of metadata for the object, it may be presented as default values.

When a document is assigned an ID, agent objects representing higher-level abstractions may be implicitly created in the archive, unless already known. Optionally, metadata may be entered for these objects, and they may be assigned IDs then or later on.

Specific Document. (*Item*) The ID applies to a snapshot of the network document, i.e., its exact content, down to the bit level, at one specific point in time. Retrieval conditions, such as cookie values, query parameters, timestamps etc. are part of the definition. When the ID is assigned, a copy of the document is stored in the National Library's archives. This copy may later be used for content authentication purposes.

Dynamic Document. (*Manifestation*) The ID applies to a network document that may change over time. As for Specific Documents, the registrant may modify the default selection of components, cookie values, query parameters etc.

Conceptually, a Dynamic Document encompasses the entire revision history of its content. The National Library of Norway may take a snapshot of the content (which, incidentally, is a Specific Document) at the time of registration and at later intervals; there is no guarantee that the entire history is saved. The ID, associated with one URL, spans the document's lifetime. This starts, by default, at registration time and extends until explicitly terminated. If the document is moved, the registrant may modify the URL in the definition. This change becomes part of the document history.

Expression. An *Expression* is abstract, not in itself manifest on the network. An agent object is created (unless already existing) in the National Library archive. The assigned ID is associated with this object, and the registrant usually supplies metadata to be associated with it. The object may have relationships to other objects, both Expressions and other entities such as authors and translators. It may have a history, and this would relate to digital and/or physical instantiations at Manifestation level.

Work. Like an Expression, a *Work* is abstract, represented by an agent object. An agent object is created when the ID is assigned. Like an Expression object, a Work object relates to other entities, such as authors, and to instantiations like Expressions. A registrant requesting an ID for a Work will usually supply metadata describing it.

3.2 Document Extent

Static/Dynamic Document: An ID allocation service may e.g., suggest a set of components (URLs to text, images etc.) as the identifier extent. The registrant may remove the URLs he considers not to belong to the document, and he may add other URLs, e.g., tie several web pages together as one document. Browsing facilities will be available to the registrant.

Static Documents cannot include dynamic/streaming information, such as web camera images, as these cannot be downloaded as a well-defined unit. Cookie values, browser ID and URL query parts are by default associated with the ID, but may be edited by the registrant before submission. Each component (URL) of the document

may have cookies and query parts. URLs and registration time are parts of the identifier definition.

Expression/Work: As these are abstract concepts, materialized on the net only through Dynamic or Static Documents, the scope cannot be identified as a set of specific components. The main purpose of the agent object is to represent the history, and for managing relations to other objects such as authors, publishers and instantiations at lower abstraction levels (including more traditional forms of Manifestations of Expressions, identified e.g., by ISBN). A registrant may later add new lower level instantiations as they are created.

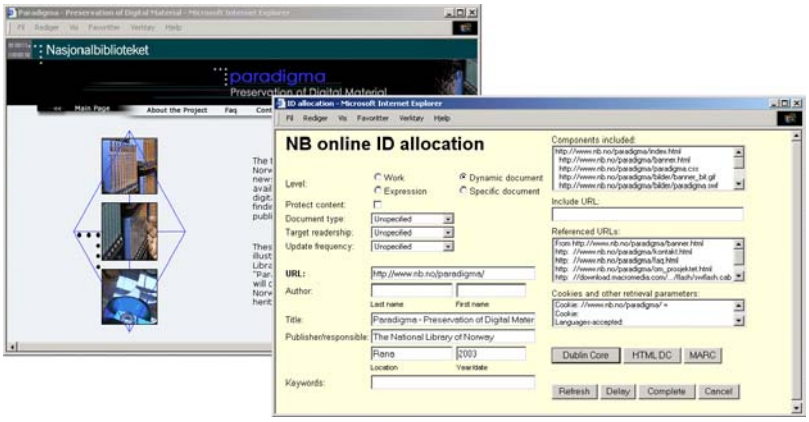

Fig. 1. Sample Web form for ID assignment. Document text is displayed in a separate window.

3.3 Metadata Registration

The registrant may supply basic metadata for the document, including document abstractions such as Work or Expression objects, which is to be identified. Any existing metadata, or information that can be extracted from the document itself, is proposed as default values. Various user groups, e.g., a publishing house or author, may be shown metadata schemes of different complexity. After entering/editing metadata, the registrant may request a text representation of the data as a bibliographic record. This can be inserted into the document before the registration is finalized.

All metadata will be reviewed/edited by librarians before it is accepted into a bibliographic database at the National Library.

3.4 Point/Fragment References

Allocating a point/fragment ID creates a *reference* object. The ID is suitable for printing e.g., in a list of references, as a citation reference etc.

A reference object is based on a defined document ID, adding the definition of a point or fragment. The user interface may offer drag&drop facilities for this purpose. Several definitions are permitted, each for a different document format: E.g., for the HTML version of the document, an URL anchor may be specified, while a page number is specified for the PDF document version.

3.5 Document Authentication

A copy of a Specific Document can be authenticated without retrieving the copy stored in the National Library's repository: This can be done by concatenating all components as octet streams without separators or padding in the order specified in the document definition, and calculating the MD5 checksum according to RFC 1321. The MD5 for the registered ID is available at any time from the online allocation service. A locally calculated MD5 value matching the value provided by the National Library may be considered proof that the copy is identical to the document that was assigned the specified ID.

A local authentication gives no indication of differences between the authoritative copy and the candidate copy. The candidate copy may be submitted to an online authentication service, and a comparison summary is returned. If the documents are not identical, the percentage of differing text lines, images and other component types may be reported. The percentage of added or subtracted materials may also be reported. The comparison is highly format specific; for formats not known by the authentication software, a component-by-component equality test may be done.

The content of the archived copy and the specific values of differences are not reported; authentication of confidential documents is therefore possible.

3.6 Relationship to Other Identifier Schemes

Network documents may be identified in other schemes, such as ISBN, ISSN, DOI and others. Paradigma sees no conflicts between these and the scheme suggested in this paper. A document may be assigned both, say, an ISBN and an URN, as alternate search keys.

URN is applicable to a larger object set; an ISBN should usually be assigned to a Specific Document only. An ISSN identifies not one specific document, but a dynamic aggregate of Specific Documents. The specification of DOI, and many other schemes, does not address several of the issues handled in our proposal. We may see a diverse use for DOI and other schemes, assuming varying semantics for properties outside the definition.

We see no reason for requiring a URN value if IDs from other schemes give sufficient identification of the object. In cases where there is a need for an ID with exactly defined semantics, an URN may be allocated in addition to another ID or as the sole external ID for the document.

3.7 Conclusions and Implementation

The Paradigma Project's main objective is to establish procedures for legal deposit of digital documents, entering these into the digital long-time storage vault at the National Library. Suggesting revisions to the current Identifier Allocation Service is part of this work. The proposed design presented in this chapter will be discussed internally at the National Library. The result of these discussions will be evaluated by the Project's steering committee, which will make the final decision regarding implementation. Paradigma has financial funding through 2004. A scheme capable of identifying all digital objects harvested from the Internet is a fundamental requirement for the Project to reach its goals. Complete definition of an ID scheme, followed by implementation of the required parts for the digital document archive to become operative, is integral to the project activity and will be completed within the timeframe of the Paradigma project.

A revised Identifier Allocation Service may be implemented within or outside Paradigma. This service and the digital document archive have several essential connection points, and if manpower resources allow, we see advantages in extending the service in parallel with Paradigma implementation. Parts of the design may be left unimplemented if there is a lack of resources; e.g. the first version may lack an authentication service, the metadata registration functions may be limited, and point/fragment reference support may be deferred.

References

1. The Paradigma Project. – http://www.nb.no/paradigma/eng_index.html (last visited July 20, 2003)
2. Norge. [Pliktavleveringsloven (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.
3. RFC 1738 Uniform Resource Locators (URL). T. Berners-Lee, L. Masinter, M. McCahill, December 1994. http://www.ietf.org/rfc/rfc1738.txt (last visited July 20, 2003)
4. International DOI Foundation home page: http://www.doi.org (last visited July 20, 2003)
5. Persistent URL Home Page: http://purl.org (last visited July 20, 2003)
6. RFC 2141 URN Syntax. R. Moats, May 1997. http://www.ietf.org/rfc/rfc2141.txt (last visited July 20, 2003)
7. IFLA Study Group on the functional requirements for bibliographic records. *Functional requirements for bibliographic records:* final report. Munich, Germany : K. G. Saur, 1998. http://www.ifla.org/VII/s13/frbr/frbr.pdf (last visited July 20, 2003)
9. ANSI/NISO Z39.56-1996 Serial Item and Contribution Identifier http://sunsite.berkeley.edu/SICI/version2.html (last visited July 20, 2003)
10. Publisher Item Identifier as a means of document identification. http://www.elsevier.nl/inca/homepage/about/pii/ (last visited July 20, 2003)
11. RFC 1321 The MD5 Message-Digest Algorithm. R. Rivest, April 1992. http://www.ietf.org/rfc/rfc1321.txt (last visited July 20, 2003)
12. RFC 3188 Using National Bibliography Numbers as Uniform Resource Names. J. Hakala, October 2001. http://www.ietf.org/rfc/rfc3188.txt (last visited July 20, 2003)