Principles of identification: European perspectives

by

Juha Hakala
Director, Information Technology
Helsinki University Library

Introduction

Identifiers, like metadata, serve many different purposes. They are of course very useful for searching, since they may provide a unique access key to the identified resource. Selling electronic and even traditional books and journals is very dependent in identifiers; if a book purchase order or ILL request contains ISBN, it is much easier to deliver the correct product. Identifiers are also good for libraries' housekeeping functions, such as deleting duplicate records from union catalogues. Identifiers are also essential for resolution services, systems which provide persistent linking between references and the resources themselves.

This article provides an overview of the present status and future challenges in identifying electronic resources, and a European view – or perhaps Finnish - on these issues. As of this writing (November 2001) there are still many open issues, and one of my aims to increase awareness of these. It is important to understand that the in the foreseeable future the way we identify resources in general and electronic resources in particular will change in a fundamental way, and this will have a major impact on our systems and the staff.

I will concentrate on the following three areas in which the emergence of the Internet has made it necessary to improve the existing infrastructure:

- Identifiers, especially ISBN, ISSN, national bibliography number and the SICI (Serial Item and Contribution Identifier), which is used for e.g. identification of articles.
- Emerging identifiers for works, and especially ISTC (International Standard Textual Work Code), which reached the ISO Committee Draft status in October 2001.
- URN resolution service, the Internet standard solution for linking e.g. resource descriptions and the resources themselves together

I will also refer in several occasions to some metadata formats such as MARC, since metadata is essential for the identification of resources and the operation of resolution services. No systematic comparison of current models is provided since that would require a lot of time. A recent survey has been done in (Snijder 2001).

ISBN and ISSN were developed about thirty years ago. At that time it was assumed that these identifiers and the related user documentation will be sufficient for a very long time. Up to the late 90's this belief was valid, but the recent technical development and especially the emergence of the Web and e-commerce have changed the situation in a few years. Now it is obvious that systems designed in the early '70s for printed publications are not suitable for the Internet usage in their present form. Either the syntax of a traditional identifier system or the rules governing its usage (or both) must be modified. And we will need completely new identifiers and formats in order to deal with new kinds of resources being published in the Web.

Identifiers – the big picture

Before the Internet, it was only necessary to identify the actual printed items, or to put it in another way, manifestations of works. However, electronic publishing requires multi-

layered identification, starting from the authors themselves, and ending in the smallest units available (for sale) in the Internet as separate items, such as a journal article or an image published in a book.

In the network environment, at least the following categories of identifiers will be needed:

- Author identifiers. International Standard Authority Data Number (ISADN) will identify each author uniquely. This is very important when there are many "legal" forms of the author's name due to e.g. transliteration. Supporting author searches in virtual union catalogues will be much easier if ISADN can be used for bringing the different name forms together. On the other hand, collecting societies need ISADN for being able to pass copyright payments to the correct person.
- Identifiers for works. These will be needed since each work may have a large number of different manifestations, and something is needed to bring these manifestations together. ISO is currently developing a family of International Standard Work Codes. Initially there will be three standards: International Standard Audiovisual Number (ISAN) for audiovisual materials, International Standard Musical Work Code (ISWC) for music works, and International Standard Textual Work Code (ISTC) for textual materials. There will also be an ISWC standard for still images, but its development has not begun yet. Systems supporting these identifiers should also support the IFLA Functional Requirements for Bibliographic Records; although some experimental systems have already emerged, it will take years until most integrated library systems can deal with both works and manifestations.
- Identifiers for manifestations of works, such as book editions. This is the category, which is already familiar to us; all traditional identifiers such as ISBN and ISSN belong to this class of identifiers.
- Identifiers for contributions (component parts) within manifestations. There are two emerging identifier systems in this group: Serial Item and Contribution Identifier (SICI) for articles, and Book Item and Component Identifier (BICI) for e.g. book chapters. Unfortunately it seems that BICI will at least for the time being be not widely implemented.

Managing all these systems will require a lot of work. All identifier systems demand some underlying metadata in order to work well; for instance an ISSN would be worthless without metadata describing the serial to which the identifier has been given. In fact, the ISSN standard requires that each serial that has received an ISSN must be catalogued. No such requirement exists for ISBN, for the time being: the new ISBN draft does require delivery of ONIX metadata.

Also identifiers for works and component parts need metadata, even if only the identifiers for works will make metadata obligatory. But who is going to catalogue for instance all textual works such as articles and books available in the Internet? There is no final solution to this question, but co-operation between authors, publishers and libraries will help in covering the most relevant material. New technologies will assist us; for instance, if XML is used as the production format, it is possible to embed sufficient metadata into the resource, and extract this information from it later on.

Traditional identifiers are already being used for identification of electronic resources. Thousands of e-journals have an ISSN, and many thousands of e-books and CD ROM disks have an ISBN. But this does not mean that ISSN, ISBN or other identifiers are really ready to deal with electronic resources. At least the following problems may emerge:

- An identifier system does not necessarily scale up so that each document belonging to its scope ("serials", "books", and so on) could be dealt with. ISBN was designed for

printed publications; can it deal with both printed and electronic books? The answer is no.

- If the identifier assignment and metadata creation can't be automated, extending the identifier provision from printed materials to electronic resources will require a lot of additional staff. Some identifiers have very complex syntax, which makes providing them manually very difficult.
- It may be unclear how an identifier system can be applied to the electronic resources because the user guidelines of the system cover only printed resources. Extension of the rules to apply also to electronic materials may be hard, for instance because the electronic resources are changing all the time. Aiming at a moving target is seldom an easy task.

Whether these problems have been solved already can be seen from the next chapters.

ISBN

International Standard Book Number was developed in the late 60's. The system has been very successful; in the year 2000 there were 152 countries using ISBNs. As a rule the ISBN centres function well, although some countries suffer from staff limitations.

ISBN consists at present of four parts: country identifier, publisher identifier, publication identifier and control character. Country identifier may refer to a single country (951 = Finland), or a language area (3 = Germany, Austria and German speaking part of Switzerland). Given this structure, an ISBN contains a good hint as regards the geographical location of the publisher and a (national bibliography) database, which may contain information about the book (there is no obligation to catalogue every book which gets an ISBN).

From the ISBN point of view, physical format of a publication is not important. Any book, printed or electronic, should receive an ISBN. Therefore using ISBNs in the Internet should be easy. Unfortunately there are at least two major problems.

In the Internet, basically anybody can be a publisher. Because ISBN identifies also the publisher, the demand for publisher identifiers grows exponentially. To some extent ISBN can cope with this, since each country usually has dedicated a publisher identifier (in Finland 952-91) for books published by individual people. But this mechanism is not flexible enough for Web publishers. So, there is a need to extend ISBN in order to accommodate a very high number of publisher identifiers.

The Internet has increased the number of published books, both on work and especially on manifestation level. It can be assumed that a large percentage of printed books will in the future be published also in digital form, either at once or retrospectively via digitisation. There may be multiple parallel digital variants of a book, and any book may consist of many component parts; for instance, each chapter may be an independent resource available for buyers as a separate item.

The ISBN system, having born in the 60's, was designed for traditional book publishing. There would have been enough numbers in the ISBN for printed books for quite a long time. With the additional load from electronic publishing it is possible and even likely that the ISBNs will run out by 2010. Early adoption of BICI, which at present seems very unlikely, would give the traditional, 10-digit ISBN a little bit longer life expectancy.

In order to avoid ISBN shortage, the ISBN system has to be extended so that there will be enough numbers for years to come. And the extension must be decided upon quickly, since otherwise the library system vendors will not have time to modernize their applications in

time. It must be kept in mind that since ISBN data is stored in many places in the integrated library systems, any change involving ISBN may be hard to implement.

The ISO has published in November 2001 a working draft for the new ISBN (ISO/WD 2108). It proposes that ISBN should be extended from 10 to 13 numbers. This would be done by adding a new element, product prefix (in practice the EAN book code "978") into the beginning of the ISBN. This would in practice double the capacity of the ISBN system.

The main merit of the proposed new ISBN would be, in addition to enhanced capacity, interoperability with the EAN system, which allows a maximum identifier length of 13 characters. Given the vital role ISBN plays in e-commerce, it is unlikely that the new ISBN could be independent of EAN. Otherwise ISBN could be made all at once for instance 16 bytes long, which would definitely satisfy the needs of the ISBN-thirsty for a long time.

In order to "squeeze" more identifiers out of the EAN-compliant ISBN, it has also been proposed that the new ISBN should be an ISSN-like dumb number. This would, of course, add the capacity of the system a lot, but many ISBN centres especially in Europe were opposed to this idea. The fact that ISBN is an intelligent identifier, which shows the publishing country or region and the publisher is seen as an important part of the system. This feature is indeed important not only for "ideological" reasons, but because resolution of ISBNs – that is, retrieval of bibliographic information and/or the resource itself using the identifier as the starting point - can be accomplished with the present ISBN and national bibliography databases. Alas, dumb ISBN can't support the resolution process.

Resolution of ISSNs within the URN system is only possible because of the ISSN database maintained by the ISSN International Centre in Paris. For ISBN, no such global database exists, and building one would be very difficult.

The proponents of the dumb ISBN point out that some international publishers occasionally use wrong country codes; for instance a publisher with headquarters in Germany may acquire an ISBN with the country code "3" for a book published in the U.S. However, such a book would be deposited in Germany and catalogued into the German national bibliography, which is where the URN resolution process would go to look for the data.

At the moment it seems likely that the new ISBN will have 13 digits. Given the general opinion of the community is quite sure that the new ISBN will not be dumb and it will not be a hexadecimal either. The ISBN community is pressed for time; the new identifier should be in use in January 1, 2005. This is only possible if the new ISBN is agreed on at least 1-2 years earlier, in order to allow the library system vendors to modify their systems in time.

As of yet it is too early to say when the new ISBN will be approved by ISO, even if the ISBN community would be able to deliver more finished drafts of the standard quickly. Some difficult issues remain; for instance, it is not clear that the organisations applying ISBNs will be able and willing to provide mandatory bibliographic description of the book in ONIX format into the global database maintained by the ISBN agency or its designated registration service.

ISSN

International Standard Serial Numbers are widely used for identification of serials, such as journals, newspapers, periodicals and so on. Contrary to the ISBN, ISSNs are dumb, that is, they do not give any hint as regards where and by whom the journal is published. Luckily the ISSN International Centre, which co-ordinates the usage of ISSN, maintains a global ISSN database which in Spring 2001 contains about one million records. Every ISSN

allocated must be accompanied with metadata, which must be sent to the ISSN International Centre for loading into the global ISSN database.

The syntax of the ISSN, eight digits of which the last one is check digit, allows for 10 million ISSNs. Since only one million ISSNs have been used by now, the identifier will scale up to cover electronic journals quite well. Up to now the average annual rate of ISSNs assigned is about 50.000. Although the ISSN guidelines require separate ISSNs to be assigned to the printed and electronic versions of a journal, this has not added the consumption of ISSNs significantly – yet.

The ISSN community has done a lot of ground-breaking work in defining electronic serials. An electronic journal does not need to be issued in volumes and issues; indeed, any Web site under which new texts are collected could in theory be regarded as a serial or at least some kind of continuing publication. Because of this change in publishing, cataloguing rules have already been modernised; new ISBD for continuing resources is available. Also the Anglo-American Cataloguing Rules have been modified; alignment between ISBD and AACR has not been a trivial task, but as of this writing these cataloguing rules do more or less agree on what to do with serials. Whether the compromise is workable will be seen; the main problem may be that electronic serials change all the time, which makes necessary constant editing of the related bibliographic data.

Unfortunately the existence of revised rules and user guidelines do not necessarily solve the whole problem. Internet journals are also difficult to deal with because of their tendency to disappear entirely or change location (URL). Simple and efficient means for archiving the electronic journals should accompany cataloguing efforts. Web archiving projects such as the Nordic Web Archive (http://nwa.nb.no) initiated by the Nordic National Libraries, or the Internet Archive (http://www.archive.org/), will provide a partial solution to this problem.

As the printed and electronic serials will co-exist for a long time, national and regional ISSN centres must deliver more ISSNs and create more bibliographic records than in the past. It is evident that more staff will be needed in the national centres and the international centre; for instance in Finland one more full-time cataloguer for electronic serials was needed.

In order to streamline the serials cataloguing the International centre must investigate the possibility of revising its technical infrastructure so that direct on-line copying of serials bibliographic records from the ISSN database will be possible. The ISSN system will also play a crucial role in resolving URNs based on ISSN or SICI, as will be seen later.

Compared with traditional publishing, Internet publishing has much bigger granularity. Instead of accessing or purchasing a book or a serial volume/issue, a user may access or buy a chapter in a book or a single article, if only there is metadata, which helps him/her in locating the relevant information. It must be admitted that as a rule the libraries have not been capable of dealing with articles too well, mainly because the volume of data is too big. Now, with the Internet allowing direct access to the articles, this problem is much more acute than before.

Publishing community is reacting to this change in several ways. As far as identifiers are concerned, development of identifiers for component parts (such as articles and book chapters) and works (such as Shakespeare's Hamlet) are of major importance.

SICI & BICI

Serial items and contributions such as articles can be identified with Serial Item and Contribution Identifier, or SICI. Unfortunately only a few serial publishers are actually using

SICI; one reason for this lack of interest may be that SICI is still only an American national standard (ANSI/NISO Z39.56), although the first version of SICI was published in 1991. As of this writing there are no plans to standardise SICI in ISO. As of this writing there are discussions in NISO about revising SICI; the decision has not been made yet.

Another reason for the publishers' lack of interest is the complexity of SICI. It consists of the serial's ISSN, data about the issue (Item segment), article/contribution data (Contribution segment) and control segment. Mark Needleman's article "Computing resources for an online catalog – 10 years later", published in Information technology and libraries, volume 11, issue 2 (June 1992), page 168-, gets the following SICI:

0730-9295(199206)11:2<168:CRFAOC>2.0.TX;2-#

For a layman the above string of characters does not make much sense. A librarian can work out everything else but the control segment, which says that this SICI is based on the version 2.0 of the standard and that the identified article is a printed text. Had it been a network document the code TX would have been replaced with code CO.

Anyway, creating SICIs is a major problem as long as they can't be generated automatically from the articles or article metadata. Automated creation of SICIs has been taken into account in the design of SICI, and there are a few pilot systems capable of this task.

In the other end of the SICI utilization chain, using SICIs is very hard unless they can be used as click-able hypertext links. Very few people could memorize SICIs and typing them for instance into an OPAC is not pleasant either.

However, in the present networked environment there is no reason why SICI would not become a very popular building tool for linking article references and the referred articles together as a part of a larger resolution service. Because of its rather versatile structure, SICI is very scaleable; it is capable of identifying millions if not billions of articles. In this respect the only, and luckily a minor, problem of the present SICI is that in some occasions different electronic versions of an article may get the same SICI. As of this writing, the ISSN user guide says that the original printed serial and its digitized copies will get the same ISSN. If we have an article, which is scanned with 600 dpi resolution for printing purposes and 75 dpi for display, both versions will get the same SICI. The intellectual content of the document is the same in both cases (provided that all details have been captured with the lower resolution) but the usage of the document is quite different, and therefore it would be important to make a distinction between these two variants.

However, I do not believe that the technical problems have been the main reason for the slow adoption of SICIs. The problem may well be primarily an organisational one. Contrary to ISBN and ISSN, there is no international centre, which would co-ordinate the development and usage of SICIs, and no dedicated national and regional centres taking care of the promotion of the system in their own areas. Publishers are supposed to generate SICIs by themselves, on the basis of the ISSN number. Unfortunately, as there is no supporting organization, few publishers are aware of the existence of the SICI. Those who know about it do not get help in implementing SICI-based services.

National ISSN centres should inform serial and newspaper publishers about SICI and investigate possibilities for implementing them. In the long run, much of the scientific and newspaper publishing will be Web driven, and the national library will receive the articles via voluntary or legal deposit. If the articles contain SICI, or the document structure – for instance, XML document type definition developed for newspaper articles – enables automated generation of SICIs, the national library will have a good starting point for

managing electronic deposit for these articles. It is generally agreed that identifier is one of the most important elements of preservation metadata. So, if the articles do not contain SICIs and the identifiers can't be constructed from elements in the articles, the national library should create the SICIs manually – which is probably impossible – or use other identifier, such as national bibliography number.

Unlike ISO standards, all ANSI/NISO standards are available for free in the Internet. The address is http://www.techstreet.com/nisogate.html. The easy availability of SICI text does simplify a little bit the task of informing the publishers. Moreover, it does not make any sense to give ISSNs to articles. Retrieving all articles ever published in the electronic version of the Time magazine with its ISSN would be, as the Americans say, counterproductive. The publishers will eventually identify electronic articles with SICI or internal identifiers embedded into for example DOI resolution service. The libraries should convince them that the SICI-based choice is the correct one.

BICI, or Book Item and Component Identifier, looks a lot like SICI. However, BICI is much more in danger than SICI. The first reason for this is that BICI is not yet a finished ANSI/NISO standard but only a draft until January 2002. What happens when the review period can't be known yet. BICI-based experiments have, as far as I know, not been popular among software developers.

If BICI is not widely implemented in the near future, demand for ISBN numbers will grow a lot, since the publishers intend to start selling e-books and their component part. Although e-books seem to be a rather dormant business at the moment it is not safe to assume that this state of affairs will continue. In the Internet best practices and business models may change quickly.

It seems at the moment that ISBN numbers will be used instead of BICIs for identifying component parts of books. Accommodating BICI would require thorough revision of the IT systems used in book trade, whereas extended ISBN usage is possible with present applications (except that the systems need to be able to deal with the new ISBN, with 13 digits).

Libraries' possibilities for fostering usage of BICI are at present limited. The least we can do is to study the standard ourselves in order to find out what it can do for us, and for publishers. It is also important to give feedback to ANSI/NISO about the BICI, and point out that even if it is not used yet, there is a great demand for being able to identify the contents of books in more detail than ISBN can or should accomplish.

Identification of textual works

ISTC, or International Standard Textual Work Code, is an ISO standard currently under construction. The ISO committee draft – the earliest public version of the text – was released in autumn 2001. The following description of the standard is based on this version of the text.

The purpose of the ISTC is to enable the efficient identification of textual works. ISTCs shall not be applied to manifestations of a textual work; other identifiers (ISBN, BICI, ISSN and SICI) already exist for this purpose. Thus the original version of the Finnish national epic, Kalevala, will get just one ISTC, although its printed and electronic versions have multiple ISBNs.

An ISTC shall consist of 16 hexadecimal digits using numerals 0-9 and letters A-F. It shall be made of the following parts:

- Registration agency element

- Year element
- Work element
- Check digit

Example: ISTC 0A9-2002-12B4A105-6

There will be room for 4096 Registration agencies. The ISTC Registration Authority will supervise these agencies, and promote & co-ordinate the ISTC system.

Each agency will be able to deliver annually a billion ISTCs up to year 9999. The committee (and any sane person) believes that this amount of numbers will be sufficient even in the very distant future.

The ISTC committee has had some preliminary discussions about quality criteria for the registration agencies. Everybody agrees that proven capability to create metadata for works (or manifestations) is an important criterion. Thus national libraries are among the strong candidates for the job. But how difficult will it be?

Kalevala illustrates well the main problem in developing an identifier for works: how to define a work. Is the Kalevala in English the same work than the Finnish original? Is the Kalevala illustrated by Akseli Gallen-Kallela a different work than the first edition, which lacked pictures? How about the abridged version of Kalevala, intended for kids, written by Elias Lönnrot himself back in 1850's?

If the ISTC working group had consisted of library experts only, we could have used the terminology and principles defined in IFLA Functional Requirements for Bibliographic Records and Anglo-American Cataloguing Rules (for a good overview of these, see Tillett, 2001).

IFLA FRBR study defined the entities work, expression, manifestation and item. A work, such as Kalevala, may be realized through one or more than one expression, which may be embodied in one or more than one manifestation, which may be exemplified in one or more than one item.

So, Kalevala is a work, which is expressed for instance in Finnish and in Czech translation; the latter has been manifested in one or more than one manifestations (first edition, subsequent editions), and Czech libraries and book lovers have items of these manifestations at their bookshelves.

Terms work, manifestation and item are familiar and somehow intuitive. But expression was in a way "invented" in the IFLA study. The borderline between works and expressions is not necessarily clear, since both works and expressions are intellectual or artistic creations of mind. For instance translations of a novel or performances of a composition are expressions, but this is obvious only for a reader with library background. And for non-librarians it will be difficult to approve of the detailed analysis our cataloguing rules make regarding where to draw a limit between two works. For instance, a faithful translation is only an expression, but a free translation is a work. This means that for instance every translation of Joyce's Finnegan's wake is destined to be a new work, since there is no way to make a literal translation of Joyce's late masterpiece.

Other communities within the book trade have not approved of the concept of expression. For instance the model developed by the INDECS project (Interoperability of Data for Electronic Commerce Systems) does not make the distinction between an abstract notion such as Verdi's Requiem, and the different realizations of it. Actually, to make things

worse, INDECS model does contain the term expression, but it means events that are creations (works) in themselves, such as a performance of Verdi's Requiem.

In the committee developing the ISTC standard there are people from many communities, including INDECS and library worlds. In the committee it was impossible to talk meaningfully about expressions, since the word had at least two different meanings, depending on who was talking. It was also impossible to adopt any existing model for defining work as such. But nevertheless it was possible to produce a document, which according to my opinion is not in conflict with IFLA FRBR or other popular and widely adopted models.

All parties in the ISTC committee agreed that the ISTC should be given to all distinct creations of mind, that is, to works and expressions (in IFLA meaning of these words). Thus every English translation of Kalevala will get its own ISTC. There will also be metadata related every ISTC; one of the most important metadata elements being links to other expressions of the same work and possibly also other works which belong to the same family of works (such as all performances of Verdi's Requiem, or all translations of Kalevala). There are of course problematic areas; for instance compilations are sometimes works in their own right, but they may also be just collections of existing works which do not deserve a new ISTC.

Details concerning the definition of work will not be written into the standard or even its annexes, but into a user guide, which can be changed more easily than the standard itself. This will help in adapting the text according to the user needs. The aim is to avoid specifying very strictly how ISTC is to be used, since this might prevent usage that would make perfect sense a few years from now.

From libraries' point of view, ISTC is a big challenge. Every ISTC must be accompanied with metadata, and since ISTC will be one of the basic building blocks in the future e-commerce systems, there will be great demand for ISTCs and related metadata.

However, our integrated library systems currently support only description of manifestations. Only a few vendors have already begun to develop systems, which support cataloguing of works. But doing this is probably trivial compared with the monumental task of creating metadata for works. Nobody knows for sure yet if it will be possible to generate descriptions of works from our existing bibliographic records with sufficient accuracy. If not, large-scale retrospective cataloguing of works would be very slow because of staff limitations; the process would proceed gradually and take decades.

To make things even more complicated, there are other emerging ISO identifier standards for works. These are ISAN (International Standard Audiovisual Number) and ISWC (International Standard Musical Work Code). None of these systems is as of yet finished, but ISAN is already a Final Draft International Standard (FDIS), while ISWC has reached DIS status. Although ISAN, ISWC and ISTC are being prepared more or less at the same time, the committees consist of different people with different backgrounds.

In practice this has meant that the standards, at least during the preparation stage, have not been fully aligned, functionally or conceptually. It may be argued that this is not a problem, but if the final versions remain inconsistent this may cause some functional issues. For instance, all work identifiers will be 16 digits long, but the components will be different. In one system version number (of the resource) was part of the identifier. At least this author feels quite strongly that version information should be included into the metadata, not into the identifier itself.

In order to avoid possible problems ISO has launched a task force, which will analyse all ISO identifier standards in order to see what kind of alignment between these standards would be desirable and possible. This kind of co-ordination is very important when the number of identifier systems is growing fast. And it may well be that a single organization such as the national library or a large publisher must deal with many of the systems simultaneously.

Given the huge amount of metadata that the new identifier systems require, it is obvious that libraries, publishers and authors themselves must join forces in getting the job done. Articles will not be catalogued exhaustively unless the people who write or publish them help librarians to do the job. Establishing this cross-organizational co-operation will be an interesting challenge, both technically and mentally. Librarians have for a long time regarded cataloguing as their own domain. Letting “amateurs” to do the job and not revising it afterwards may be difficult for some colleagues.

Resolution services

For hand-held materials finding a resource was a simple task, provided that the book or journal issue was filed correctly. A customer checked the call number from the catalogue card or OPAC record, walked to the correct shelf location and fetched the thing, or made an order for a resource shelved in closed stacks.

For electronic resources things are both more complicated and easier. Systems used for linking the resource and its description or two resources directly are often called resolution services. Development of these services is as of this writing a rapidly advancing area in the Internet.

Resolution services can be roughly divided into static and dynamic. Uniform Resource Locators and HTTP protocol are ingredients of a well-known static service; the one used in the Web. The URL-driven linking works fine as long as the resource is not modified, stays in the same location and the user is authorized to get it. Once the location changes, there is no easy way of finding the resource from the Web again, unless the Domain Name Service has been redirected to the new location. And there is no way to find out with URL only if the content has changed or remained the same.

Thus we have two requirements for dynamic – and efficient – resolution service for electronic resources. It must be able to adapt to the location (URL) changes, and it has to be able to take into account the user’s privileges, based on his/her credentials and / or location. A good system should also be capable of personalized information services, based on the user’s interest profiles. Last, but not least, the system should be well integrated into the Internet infrastructure.

Uniform Resource Names

Internet Engineering Task Force (http://www.ietf.org) launched the URN project soon after the Web became popular after the introduction of Mosaic, a Windows-based Web interface. In 1994 the project published generic requirements for URNs in RFC 1737:

- Global scope: the name is not bound to a location, it’s meaning is the same everywhere
- Uniqueness: the same name may not be given to two different resources
- Persistence: the name must remain the same to infinity
- Scalability: the name can be assigned to any possible resource
- Legacy support: the name must be able to support legacy naming conventions, if the other requirements can be applied to those conventions
- Extensibility: it must be possible to expand the naming scheme

- Independence: the organization responsible of the names must be totally independent

Fulfilling these turned out to be not a trivial issue; the first IETF URN working group broke up without achieving its goals.

The work was restarted in 1996, and the second working group has been more successful; in autumn 2001 the work is practically done. It has taken much more time than anticipated, but there have been good reasons for this delay, as will be explained later.

The URN working group (http://www.ietf.org/html.charters/urn-charter.html) defined first the syntax of URN. According to the RFC 2141 (http://www.ietf.org/rfc/rfc2141.txt), URN consist of three parts:

- Character string "urn:". This prefix is needed in order to make it possible to locate and index URNs from non-structured Internet resources in which there is no way to indicate the presence of URN (by e.g. using Dublin Core Identifier element)
- Namespace Identifier (NID), which identifies uniquely the identifier system used
- Namespace specific string (NSS), which contains the actual identifier

Each NID has to be registered. This process and the data applicants need to provide is defined in RFC 2611 (http://www.ietf.org/rfc/rfc2611.txt). ISSN International Center has registered NID "ISSN" for International Standard Serial Number (see http://www.ietf.org/rfc/rfc3044.txt). Registrations for International Standard Book Number (NID "ISBN") and national bibliography numbers (NID "NBN") were approved by the Internet Engineering Task Force (IETF) in October 2001 (see http://www.ietf.org/rfc/rfc3187.txt and http://www.ietf.org/rfc/rfc3188.txt). Registration request for SICI NID was sent to the IETF in summer 2001.

The process of registering these namespaces has confirmed the preliminary analysis done in RFC 2288 (http://www.ietf.org/rfc/rfc2288.txt): the most important identifiers used by libraries can be used as Uniform Resource Names. RFC2288 gave no details on how the resolution actually takes place; this information is provided in the namespace registrations. As can be guessed, the technique varies a lot depending on the identifier system.

Resolving ISSN-based URNs is easy, since there is a single resolution service, the ISSN database. Resolving ISBN-based URNs is a bit more complicated, since there is – for the time being – no global ISBN database. In order to find the correct place, the resolution process must check the country identifier, and then proceed to the correct national bibliography database or a set of them, which has been specified into the DNS system. For instance, if the ISBN begins with 951 or 952, the correct place to go is the Finnish National Bibliography database. In order to resolve SICI-based URNs, it is necessary to go first into the ISSN database, retrieve the bibliographic record describing the serial and check if the record contains a link to a database, which contains the full text of the article or bibliographic information about it.

According to the RFC 3044, an URN based on ISSN has very single syntax:

urn:issn:<issn number>; for instance urn:issn:1560-1560

Similarly, URN based on ISBN will have prefix urn:isbn: and URN based on NBN prefix urn:nbn: attached in front of the ISBN or NBN string.

The simple structure of URNs (and the fact that they are free) makes it possible to generate URNs automatically from the existing identifiers. In fact, it is not even necessary to store URNs in a database. If a request to resolve an ISBN-based URN arrives to an

OPAC, the system can remove the urn: prefix, check the NID – in this case, isbn: - remove it, and then pass the request to the ISBN index to see if the ISBN stored in the NSS part of the URN can be found. If the answer is yes, the library system can deliver the requested information – either bibliographic record, list of URLs or the electronic book itself, depending on what the patron asked for, and what access privileges she has.

By standardizing the namespace registrations the Internet community can control that the requirements of RFC 1737 are met. The registrants must describe themselves and their identifiers and the technique via which the URN resolution mechanism will be built. In the URN system one namespace – for instance the ISSN namespace – may rely on centralized service, but other namespaces such as the ISBN and NBN namespaces will be decentralized.

The services the URN system provides have been defined in RFC 2483 (http://www.ietf.org/rfc/rfc2483.txt). A user may order bibliographic information about the resource, its location (URL) or the resource itself. In order to avoid frustration, the user interface should be bright enough to hide those options, which are not available. Of course, context sensitive linking will depend on OpenURL and similar protocols and their implementations.

In the present Internet, URN resolution will be based on Domain Name Service. The idea is that a user can type URN into the Location-window of his/her Web browser or other Internet client, just like URLs can be used now. Once the user has hit the Enter key, DNS service will find the resource using the metadata stored in the DNS system. In the future, when DNS is replaced with a more advanced system, URNs will be resolved in it. Nothing in the URN syntax or the way URNs are embedded in resources will imply usage of DNS or any other resolution service. Please note that for the time being Digital Object Identifiers are stored as HTTP links, but in the future even DOIs may be used as URNs, giving them more solidity.

The Internet Assigned Names Authority is responsible of building the URN Resolver Discovery Service into the address urn.net. This DNS server will know – on the basis of registration requests and additional information provided by implementers – the location of all URN resolution services in the Internet. This information will be propagates throughout the entire DNS system via the routine DNS means. For instance, the urn.net server and many other name servers will know the address of the resolution service for ISSNs, the ISSN database. Since this information is available in the DNS there is no single point of failure in the system, and there is no technical limit to how many URN resolution services can be built for the defined namespaces.

The metadata needed for URN resolution is stored in DNS resource records, which normally link Internet names and IP numbers (such as www.helsinki.fi and 128.214.4.1) but can also deliver more complex information. The most common DNS application, BIND, supports URN resolution already.

In spite of the URN infrastructure being in practice finished, there are not that many URN services available. In the library sector, the first URN resolution service was built by the ISSN International Centre (see http://www.issn.org), but as of this writing the library system vendors have not yet built URN-based services. There are a few good reasons for this.

First, some URN standards are still unfinished. It seems that these standards will be finalised in the near future, but this has been the case for quite a while already. However, all comments from the Internet Engineering Steering Group – which reviews all URN standards – have been taken into account in Internet drafts delivered in October 2001.

Thus, if all goes well, all URN standards will be approved in early 2002. Only time will tell if this is indeed the case.

Second, the Internet Assigned Names Authority has not yet created the URN Resolver Discovery Service. This in turn means that know-how about the existing URN resolution services may at present not be available in the DNS.

Third, support for URN resolution is not built into all Web browsers. For the time being only Microsoft IE has this feature, but unfortunately the Internet Explorer 5.0 implementation of URN resolution is not fully compliant with the Internet standards. If usage of URNs required installation of a plug-in, most patrons will not bother doing it. Knowing this, libraries and other potential users have not been too busy implementing URN resolution services.

Fourth, because the URN initiative has not yet resulted into highly visible results, libraries have not requested their system vendors to implement URN resolution services. However, since the national libraries worldwide have committed themselves into using URNs, it is certain that the implementations will follow once the basic infrastructure is in place. The experience gathered up to now indicates that building a URN resolution service into a library system should not be too difficult; the ISSN International Centre was able to do it in a few weeks, using only a small number of skilled programmers.

URN implementation will not be only a technical challenge. Although URNs as such are available for free, building and maintaining the resolution services (and archiving the electronic resources which the service covers) will definitely not be free. Somebody has to pay the bill, and for the time being there are no volunteers. Nevertheless, national libraries have committed themselves to supporting the URN initiative, and in the long run this will lead into emergence of URN-based services. These services may or may not be free; the important thing is that we must use better tools than URLs for linking our bibliographic records and the electronic resources they describe.

References

Snijder, Ronald: Metadata standards and information analysis: a survey of current metadata standards and the underlying models. Electronic resource, available at http://www.geocities.com/ronaldsnijder/. Referred at 21.11.2001.

Sollins, Karen & Masinter, Larry: Functional requirements for Uniform resource names. RFC 1737. Electronic resource, available at http://www.ietf.org/rfc/rfc1737.txt.

Tillett, Barbara: Bibliographic relationships. In: C. A. Bean & R. Green: Relationships in the Organization of Knowledge, 19-35. 2001.