Journal of Documentation, Vol. 55, No. 1, January 1999

NATIONAL BIBLIOGRAPHIC RECORDS IN THE DIGITAL INFORMATION ENVIRONMENT: METADATA, LINKS AND STANDARDS

MICHAEL DAY, RACHEL HEERY *and* ANDY POWELL
{m.day; r.m.heery; a.powell}@ukoln.ac.uk

UK Office for Library and Information Networking, University of Bath
Bath BA2 7AY

This paper reviews BIBLINK, an EC funded project that is attempting to create links between national bibliographic agencies and the publishers of electronic resources. The project focuses on the flow of information, primarily in the form of metadata, between publishers and national libraries. The paper argues that in the digital information environment, the role of national bibliographic agencies will become increasingly dependent upon the generation of electronic links between publishers and other agents in the bibliographic chain. Related work carried out by the Library of Congress with regard to its Electronic CIP Program is described. The core of the paper outlines studies produced by the BIBLINK project as background to the production of a demonstrator that will attempt to establish some of these links. This research includes studies of metadata formats in use and an investigation of the potential for format conversion, including an outline of the BIBLINK Core metadata elements and comments on their potential conversion into UNIMARC. BIBLINK studies on digital identifiers and authentication are also outlined.

INTRODUCTION

National libraries, as they have developed historically, are important organisations which collect, preserve and make available publications which are seen as a major part of a particular nation's history and cultural heritage [1]. In support of these 'core' roles, nearly all national libraries have taken on the important task of managing a national bibliography, variously viewed as an official record of a nation's intellectual heritage or of its publishing output. Production of national bibliographies is closely related to the process of legal deposit, a process that varies significantly between countries worldwide. One of the most important challenges facing national bibliographic services is the increase in electronic publication as, traditionally, electronic resources have neither been covered by legal deposit legislation nor included in national bibliographies.

The BIBLINK project emerged as an initiative amongst a group of European national libraries to address the future role of national bibliographies in relation to electronic publications. In order to place BIBLINK in context we will briefly review some of the current concerns for national bibliographic services.

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 55, No. 1, January 1999

The national bibliography is typically a record of what is legally deposited, and the legacy of print culture means that in many countries legal deposit is limited to books. Legal deposit policy is now being reconsidered to take account of electronic publication with several national libraries moving towards experimental deposit of 'physical' electronic publications. The position as regards networked information is more complex, where the document may be dynamic and not easily isolated for deposit. A recent international review of the deposit of non-print material gives an account of the legal situation, and the situation in practice, in a number of European countries, the US, Canada and Australia [2]. It reveals that progress towards 'comprehensive' legal deposit is far from straightforward, involving an interaction between definition of policy, realistic implementation of that policy and the availability of technical solutions.

The nature of electronic networked information means it is no longer easy to define 'publication': the low entry cost of placing information on the World Wide Web and the ease of revision has led to a vast body of dynamic information, much of which is transitory in value. Although it may be relatively easy to achieve deposit and selection of tangible electronic artefacts such as CD-ROMs, it is far more difficult to decide what is worth collecting from the web, and whether selected material can be deposited in any meaningful way.

Increasing globalisation of information and the publishing industry prompts the question 'Which publications contribute to the cultural heritage?'. It may be possible for those countries to formulate a collection policy where linguistic patterns are well defined and the size of the publishing industry is relatively small. But it is becoming increasingly problematic in relation to networked information aimed at an international audience. Selection will be required, but the criteria for selection of electronic material are yet to be fully explored and defined.

Selection itself is not new, rather the scale and nature of the material. For printed material there have been criteria for inclusion in national bibliographies over and above deposit, for example the *British National Bibliography* (*BNB*) currently selects from material received at the Legal Deposit Office of the British Library by applying an exclusion policy whereby reports are excluded and entered into the SIGLE system, and Stationery Office publications are listed elsewhere. Various approaches to selection are emerging. Some countries, such as Norway, are collecting not only electronic artefacts but are taking snapshots of the World Wide Web in their domain area. There is acknowledgement that some selection will be required regarding what is to be preserved over time, and the technical barriers to provision of access remain. The National Library of Australia is proposing to select exemplars of various genres of electronic material such as home pages: 'It has never been possible to collect everything in print: the rapidly increasing availability of electronic materials makes it likely that some materials will not be collected at all, and others only by sample' [3].

The development of national bibliographies is connected to the concept of Universal Bibliographic Control (UBC) which, together with the International MARC Programme, is a Core Programme (UBCIM) of the International Federation of Library Associations (IFLA) [4]. The purpose of UBC has been defined as 'making use of and exchanging worldwide bibliographic records created nationally, but [which are] based on internationally accepted bibliographic

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 55, No. 1, January 1999

standards' and is based on the premise that 'cataloguers in any one country are best able to describe the publications of their country' [5, pp. 16–17]. UBC presupposed the creation of systems that could be used for the international exchange of standard bibliographical descriptions of publications such as the International Standard Bibliographic Description (ISBD) and the Universal MARC (UNIMARC) format. National bibliographies now need to encompass metadata relating to electronic resources that may contain additional content (digital preservation data, terms and conditions of use) not required for print material. New formats are being developed to contain such data and the traditional MARC formats are being revised to accommodate at least some of this additional information.

In summary, several important issues need to be resolved. Firstly, the 'core' role of the national library may need to be redefined in the light of the advent of the digital information environment and the resulting globalisation of information. Given greater globalisation, it remains to be seen whether national libraries and national bibliographies are the optimum model for ensuring access to the digital collections of the future [6]. In parallel, national libraries need to continue reassessing their priorities with regard to legal deposit legislation [7] and digital preservation [8]. Where national libraries act as national bibliographic agencies, they will have to interact with the producers and publishers of electronic information with the intention of creating systems and formulating the standards that will facilitate the flow of bibliographic metadata. In 1980, Ross Bourne commented that 'the more that libraries rely on one another, the more necessary it is that they speak the same language' [9, p. 197]. In the digital information landscape, it is becoming increasingly clear that all players in the information chain, and not just libraries, will need an adequate way of communicating information. BIBLINK is an attempt to link some of the stakeholders in the electronic publishing process and to demonstrate means of building a bibliographic link.

In the remainder of this paper we will describe the background research undertaken by the BIBLINK partners, but firstly, for comparative purposes, we shall consider briefly the related work of the Library of Congress Electronic CIP Program.

RELATED WORK: THE LIBRARY OF CONGRESS ELECTRONIC CIP PROGRAM

Cataloguing-in-Publication (CIP) is one interface between publishers and national bibliographic agencies. With a feeling that publishers were getting increasingly interested in conducting business electronically, the Library of Congress devised an Electronic CIP (E-CIP) Program in 1993 to experiment with the creation of an electronic version of the CIP process [10, p. 178]. Publishers participating in the programme would use the file transfer protocol (FTP) to submit an electronic CIP application and manuscript to the Library of Congress via the Internet. After cataloguing, a completed pre-publication bibliographic record is transmitted to the publisher where it can be inserted into the copyright page of the printed book. The University of New Mexico Press submitted the first manuscript in November 1993.

The data supplied by publishers under the E-CIP Program are processed by cataloguers at the Library of Congress using a package of utilities using Text Capture and Electronic Conversion (TCEC) techniques, later called 'On the MARC' [10, p. 179]. These utilities permit cataloguing personnel to take electronic information

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 55, No. 1, January 1999

in a variety of formats and convert these data to usable LC MARC records. A cataloguer can physically highlight information in the source information, add ISBD punctuation and then 'click' on a particular MARC tag button. The program will then create the appropriate field in LC MARC format. All of the descriptive cataloguing information in the source text can be created in this way and the completed record can be 'cut-and-pasted' into the standard LC record creation program. The utilities will also deal with the creation of name authority records and the inclusion of table of contents data in a 505 field [11]. Further experiments with the LC TCEC software proved that it could be used to build up MARC records based on OPAC data accessed via the Internet.

The E-CIP Program demonstrates that there is scope for some profitable electronic interaction between publishers and the LC regarding the CIP process. Both publishers and LC saw distinct advantages in that it removed the requirement to send large amounts of book 'front-matter' by surface mail, it speeded up the whole process considerably and it enabled cataloguers to enrich bibliographic records by the addition of table of contents notes, abstracts, etc. with little additional resource required [12]. The use of TCEC techniques by LC cataloguers was also seen as a way of speeding up the cataloguing process itself and (possibly) producing more accurate records than the re-keying of catalogue data would permit. Although E-CIP was used to 'streamline' and enhance the CIP process for books, the techniques developed could easily be adapted for use with electronic publications. However, publishers' increasing use of standardised text encoding formats – based on, for example, the Standard Generalised Markup Language (SGML) – mean that there is the potential for a more sophisticated interaction between publishers and national bibliographic agencies.

BIBLINK

The BIBLINK project [13] grew out of part of the work of an EU concerted action known as CoBRA (Computerised Bibliographic Records Action) which was established in 1993 under the aegis of the Conference of European National Libraries (CENL) with funding from the Commission of the European Communities. CoBRA was formed in 1993 to overlook the development of national bibliographic services and to identify areas suitable for research [14]. The specific terms of reference for CoBRA include fostering 'new links between organizations involved in bibliographic record creation at all stages of the publication and distribution process' and the encouragement of 'greater standardization amongst the parties involved in records supply and use' [15, pp. 158–159]. With these terms of reference in mind, the aim of the BIBLINK project was to test a demonstrator service which would involve the establishment of an electronic link between publishers of electronic material and national bibliographic agencies for the transfer of authoritative bibliographic information [16, p. 229]. The European Commission DG XIII/E-4 under the Telematics Application Programme of the Fourth Framework Programme funds the BIBLINK project, which started work in 1996 [17]. The project is led by the British Library and its partners include the Biblioteca Nacional (Spain), the Bibliothèque nationale de France, the Koninklijke Bibliotheek (The Netherlands), the Nasjonalbiblioteket (Norway), the Universitat Oberta de Catalunya (Spain) and UKOLN.

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 55, No. 1, January 1999

THE BIBLINK DEMONSTRATOR

The core deliverable of the BIBLINK project is a demonstrator service that will allow publishers to submit authoritative metadata for use by national bibliographic services. This metadata can then be enhanced by third parties and converted into national MARC formats for potential inclusion in a national bibliography. Electronic documents have been limited to those whose content makes them suitable for inclusion in a national bibliography. The publisher-types involved in the project cover a broad range of publishing activity and within the project have been broadly divided into three groups:

- traditional – where there is a background in printed publications;
- new – where there is no such background;
- 'grey' – where publishing is not the primary business.

The differences between these publisher-types can be significant not only with regard to their knowledge of the work of national libraries and bibliographic services but also with regard to the level of metadata they are able to generate. Traditional publishers are likely to have existing contacts with national libraries and bibliographic services through legal deposit or CIP and are likely to be familiar with the creation and use of metadata. New or 'grey' publishers are less likely to have these contacts or to be familiar with the use of metadata in a national bibliography. The actual format of the publications to be included in the demonstrator was not regarded to be as important as information content, but most of the publications dealt with by the project will either be off-line, typically on CD-ROM, or published on the Internet.

The specification for the BIBLINK demonstrator was only produced after the completion of a series of reports on background issues. These included studies of the metadata formats in use by publishers and national bibliographic services, the potential for conversion between these formats, the use of unique identifiers in a digital environment and authentication. These studies raised a number of important issues as well as producing recommendations to the project regarding the demonstrator itself.

METADATA AND FORMAT CONVERSION

The BIBLINK study of metadata [18] built on a major review of metadata formats which had been carried out as part of the EU Telematics for Research funded DESIRE project [19]. The BIBLINK project was particularly interested in those metadata formats produced by organisations that have had a role in the production of bibliographic information about electronic publications. Traditionally, these organisations have included publishers, libraries, national bibliographic services, trade bibliographic services (e.g. Whitaker), abstracting services, online databases, booksellers, library suppliers and subscription agents. Increasingly, however, new bibliographic agents are emerging, including authors, Internet search services, electronic text archives and electronic repositories like the pre-print archives based at Los Alamos. The result of this diversity is a wide variety of metadata formats each based on a particular user community.

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 55, No. 1, January 1999

Lorcan Dempsey comments that libraries have a longer tradition of generating and exchanging metadata in electronic form than any of the other organisations involved in the bibliographic information chain [20, 21]. The library community has in consequence developed elaborate standards for cataloguing and for the exchange of bibliographic information or metadata. The standards include the ISBD series that form the basis of the descriptive metadata in many national cataloguing rules [22] and the MARC formats which are used to encode this descriptive metadata and other cataloguing information [23]. Despite being all based on the ISO 2709 record structure, there are many different, mostly nationally based, MARC formats. The libraries involved in the BIBLINK project all use different MARC formats, and in the case of the Koninklijke Bibliotheek, a non-ISO 2709 based format (Pica+). The project was committed to producing at least two MARC formats, UNIMARC and UKMARC, but project participants would naturally want to create additional records in formats used by their national bibliographic services. Another complication was that the BIBLINK studies were being produced at a time when both UNIMARC and UKMARC were being updated to deal with electronic publications. UKMARC was in additional 'flux' due to the planned format harmonisation with USMARC and CANMARC. A revised ISBD for electronic resources which included specific guidance for online resources like Web pages, ISBD(ER), was also produced at this time [24].

In addition to the MARC formats, there are a wide variety of other bibliographic-type metadata formats in existence. In order to analyse the different formats used by publishers and other organisations in the bibliographic information chain, a typology of metadata has been built which is based upon the underlying complexity of the various formats (Figure 1). According to this typology, there is a continuum from simple metadata like that used by web search engines, through simple structured generic formats like Dublin Core to more complex formats which have structure and are specific to one particular domain or are part of a larger semantic framework. Examples of these more complex formats are the MARC formats used by libraries and formats based on the Standard Generalised Markup Language (SGML).

The three different publisher-types involved in the project potentially use a wide variety of metadata formats. Traditional publishers will be the most familiar with generating and disseminating metadata as they produce bibliographic information for supply to the book trade and for CIP. Publishers which mark up text

Band one (full text indexes)	*Band two (simple structured generic formats)*	*Band three (more complex structure, domain specific)*	*Band three (part of a larger semantic framework)*
Proprietary formats	Proprietary formats Dublin Core IAFA/Whois++ templates	FGDC MARC GILS	TEI headers ICPSR EAD CIMI

Figure 1. *Typology of metadata formats, adapted from Dempsey and Heery [25]*

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 55, No. 1, January 1999

BIBLINK data element	*Brief description*
DC.Title	Title of work
DC.Creator	Persons or organisations primarily responsible for intellectual content
DC.Subject	Subject keywords, may also contain terms from published subject headings or classification schemes
DC.Description	Description of content or abstract
DC.Publisher	Agency responsible for producing the publication
DC.Contributor	Persons or organisations responsible for content not included under DC.Creator
DC.Date	Date of publication
DC.Format	Format information
DC.Identifier	A unique identifier, e.g. ISBN, SICI or DOI
DC.Language	Language of text
DC.Rights	Terms and conditions information
BIBLINK.Checksum	Hash value or checksum computed for authentication purposes
BIBLINK.Edition	Number of edition or version
BIBLINK.Extent	The size of an item – number of files, bytes, etc.
BIBLINK.Frequency	Frequency of issue if a serial publication
BIBLINK.PlacePublication	Geographical location of publisher
BIBLINK.Price	Price
BIBLINK.SystemRequirements	System requirements

Figure 2. *The BIBLINK Core*

based on SGML Document Type Definitions (DTDs) will also typically store metadata in document headers. Examples of SGML DTDs used for metadata about serials include the Modular Application for Journals (MAJOUR) and Simplified SGML for Serial Headers (SSSH) [26]. 'New' and 'grey' publisher-types, on the other hand, would probably not have the ability (or desire) to produce metadata encoded as SGML-based document headers. For this reason the BIBLINK study concluded that it would be more realistic to consider the use of more than one metadata format within the demonstrator. It suggested the use of an extended Dublin Core element set as a BIBLINK minimum data element set and the use of SGML-based formats, and in particular SSSH, for more complex records. The data elements in the BIBLINK minimum element set, known as the BIBLINK Core (Figure 2), were identified by mapping a list of national libraries' metadata requirements to Dublin Core.

A further study looked at metadata format conversion feasibility [27]. This meant demonstrating the feasibility of converting the chosen publisher formats recommended in the metadata study to the 'target' MARC format, UNIMARC. This study also evaluated another EU Telematics Applications Programme funded

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 55, No. 1, January 1999

project, the Universal MARC Converter (UseMARCON) project, which had produced a set of software tools which could be structured to enable conversions from one ISO 2709 compatible MARC format to another [16]. The existence of UseMARCON meant that BIBLINK could concentrate on the conversions from publisher formats to a single MARC format (UNIMARC) while intra-MARC conversions could be produced using the UseMARCON tools.

The format conversion study produced mapping tables for conversions from simple (unqualified) Dublin Core and SSSH to UNIMARC. These, once revised, could become part of the formal specification for the BIBLINK demonstrator. At this initial stage, however, they could also demonstrate that suitable UNIMARC records could be created from data held in the chosen publisher formats. A hypothetical example of a conversion is included for information (Figure 3).

The mapping tables also highlighted several areas that were problematic. Perhaps the most important of these is the fact that there is currently no natural place to record a URL (or any other digital identifier) in the UNIMARC format. The Library of Congress had approved the addition of a Subfield $u (Uniform Resource Locator) to Field 856 (Electronic Location and Access) of USMARC in 1994 [28, pp. 45–54]. At the time of the BIBLINK study, however, the inclusion of an USMARC 856-type field in UNIMARC had been proposed but not officially approved. In the meantime, it is possible that digital identifiers could have been mapped either to a UNIMARC General Note field (300), which would make the return conversion difficult, or to a locally defined field in the National Use Block (9--). The use of a locally defined field had the additional advantage that other digital identifiers could be included as necessary.

Another problem is that any UNIMARC record created from the conversion process may not be valid because it is missing one or more mandatory fields. Mandatory fields include a fixed-length Record Label and General Processing Data field (100) that would be extremely difficult to generate automatically in any conversion. This may not be a serious problem because the UNIMARC records produced as part of the BIBLINK demonstrator will undergo further conversion to a number of national MARC formats and will be enhanced before being re-converted into UNIMARC and stored in the BIBLINK database.

The SSSH to UNIMARC mapping showed that there could be potential problems with granularity. MARC records, especially minimum-level CIP-type records, typically describe serials at title level (with holdings information added) while SSSH is primarily used to describe articles in serials – although the format could also record metadata about individual issues. It would be absurd if an information-rich format like SSSH were to be used just to provide details of a serial title, publisher and ISSN. In the circumstances, it would probably be better to create a UNIMARC record for each article but this might conflict with the policies of some national bibliographies.

DIGITAL IDENTIFIERS

Identifiers are used widely in traditional publishing. Most books and serials that enter a national bibliography will have an International Standard Book Number (ISBN) or an International Standard Serial Number (ISSN). In addition, publishers

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 55, No. 1, January 1999

Hypothetical BIBLINK Core record before conversion:

```
DC.Title: Taylor-Schechter Unit Home Page
DC.Creator.Organization: Cambridge University Library
DC.Subject: Taylor-Schechter Genizah Research Unit; Cairo
Genizah; Cambridge University Library, Taylor-Schechter
Genizah Collection; Hebrew Manuscripts; Arabic
Manuscripts;
DC.Subject SCHEME=LCSH: Cairo Genizah
DC.Subject SCHEME=DDC: 016.296
DC.Description: Web pages that introduce the Taylor-
Schechter Genizah Research Unit based at Cambridge
University Library. The Unit co-ordinates research work on
manuscripts (mostly in Hebrew or Arabic) originating from
the Cairo Genizah. The manuscripts were donated to the
Cambridge University Library in 1898 and form the Taylor-
Schechter Genizah Collection.
DC.Publisher: University of Cambridge
DC.Date: 19970605
DC.Format: text/html
DC.Identifier: http://www.lib.cam.ac.uk/Taylor-Schechter/
DC.Language: en-uk
BIBLINK.PlacePublication: Cambridge
```

Hypothetical UNIMARC record after conversion:

```
101 1#$aeng
200 1#$aTaylor-Schechter Unit Home Page$fCambridge
University Library
210 ##$aCambridge$cUniversity of Cambridge$d1997
330 ##$aWeb pages that introduce the Taylor-Schechter
Genizah Research Unit based at Cambridge University
Library. The Unit co-ordinates research work on
manuscripts (mostly in Hebrew or Arabic) originating from
the Cairo Genizah. The manuscripts were donated to the
Cambridge University Library in 1898 and form the Taylor-
Schechter Genizah Collection.
336 ##$atext/html
606 0#$aCairo Genizah$2lc
610 0#$aTaylor-Schechter Genizah Research Unit$aCairo
Genizah$aCambridge University Library, Taylor-Schechter
Genizah Collection$aHebrew Manuscripts$aArabic Manuscripts
676 ##$a016.296
711 02$aCambridge University Library
```

Figure 3. *Example conversion of BIBLINK Core record to UNIMARC*

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 55, No. 1, January 1999

are adopting a variety of identifiers that identify publications at a finer granularity. The most important of these are the Serial Item and Contribution Identifier (SICI), the Book Item and Contribution Identifier (BICI) and the Publisher Item Identifier (PII) used by some STM publishers [29]. National libraries use identifiers for a variety of purposes throughout the life cycle of resources. They are useful for acquisition, especially where a policy of voluntary deposit is in place, and for registration. Identifiers also form part of a bibliographic record so that they can be indexed for retrieval purposes and are visible in things like printed listings and OPAC screens so that users can distinguish between items that are otherwise similar, e.g. different editions, reprints, serials with the same title, etc.

Project BIBLINK wanted to identify existing identification schemes that met the specific requirements of national libraries with relation to the production of a national bibliography of electronic publications [30]. It was, therefore, especially interested in those identifiers that had been specifically developed for digital resources. Location identifiers – like Uniform Resource Locators (URLs) – were not suitable in themselves because of their lack of persistence. In addition, the Persistent Uniform Resource Locator (PURL) developed by the Online Computer Library Center (OCLC) was only an interim solution and for that reason was not recommended for adoption by the project. This left two major digital identifier initiatives, the Digital Object Identifier (DOI) and the Uniform Resource Name (URN).

Uniform Resource Names are an initiative of the Internet Engineering Task Force (IETF), the organisation which oversees the development of standards for the Internet [31]. The URN Working Group propose that an URN should be globally unique, persistent, scalable, extensible and should also be able to support a variety of naming policies for the assignment of identifiers, including the continued use of legacy identifiers [32]. The Digital Object Identifier was initially developed by the Association of American Publishers (AAP), using the Handle system [33], itself a URN proposal developed by the Corporation For National Research Initiatives (CNRI). The intention was to create a digital identifier system that could help form the basis of electronic commerce.

Both DOI and URN are based on the concept of a unique identifier (an alphanumeric string) that can be used to direct a user to a particular location identifier using a resolution service. In the case of the DOI, the identifier could be resolved, via the DOI Directory, to an intermediate location that rights owners (usually publishers) could generate depending upon the access-rights associated with a particular user rather than the resource location itself. This would enable rights owners to control access to their resources. The URN and DOI identifiers take broadly the same form: a registry assigned prefix (in URN, a Namespace Identifier) followed by a suffix separately assigned by a publisher or other naming authority (in URN, a Namespace Specific String). Legacy identifiers like SICIs or ISBNs can be used as URNs, although Clifford Lynch has commented that while the content of a NSS might have structure and significance to users familiar with the practices of particular naming authorities, this content has no predefined meaning within the overall URN framework [34]. In a similar way, legacy identifiers can be used as a DOI suffix but the actual DOI itself is intended to be a simple, dumb alphanumeric string. Mark Bide has suggested the inclusion

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 55, No. 1, January 1999

of an optional but fully standardised syntax to indicate which legacy identifier is being used in a particular DOI [35, p. 11].

The BIBLINK project had its own requirements for a digital identifier. Any chosen identifier should be able to include all resource-types in the project scope (i.e. both online and off-line resources). It must also be assigned by an authorised naming authority and should preferably be standard, globally unique, persistent, extensible, human-readable, transportable by commonly used Internet protocols and possible to validate. Both URN and DOI met many, but not all, of the requirements for a BIBLINK identifier. It was envisaged, for example, that the URN would not be widely used outside the Internet context and therefore would not be suitable for use with off-line publications. The DOI would be able to cover all of the resource-types in the project scope, but is not – strictly speaking – a standard. As currently implemented, the fact that a DOI need not directly link to the resource itself, but to an intermediate location like an order form, could also be a problem. The BIBLINK study suggested that the project should use the legacy identifiers ISBN, ISSN and SICI, proposed that the DOI should be used where publishers involved with the project were generating them and recommended the use of URNs within the project although it was aware that it was likely to be some time before widespread implementations of the URN would be in existence.

AUTHENTICATION

The problem of ensuring ‘authenticity’ in the digital information environment has been recognised as one of the most important issues related to the successful development of the digital information services [36]. Discussions of authentication issues in this context tend to refer to two distinct, but related, issues. Firstly there is the problem of authenticating identity, for example, checking that particular individuals or organisations have access rights to a certain resource, or confirming that a piece of data purporting to be from a particular individual or organisation is actually from that source. The development of secure technologies for controlling access in a digital environment is vitally important to the development of online commerce and has attracted much attention. Solutions usually make some use of cryptographic techniques [37, pp. 40–50]. The second authentication problem relates to resources themselves. In a digital environment, data can be easily manipulated and modified. This is, indeed, one of the advantages that digital resources have over print. For example, databases or web pages can be kept up-to-date by constant revision. On the other hand, accidental corruption or illicit modification is equally possible, and this could be done without the knowledge of the individual or organisation responsible for maintaining access to a particular resource. It could be argued, in some contexts, that the precise reason why a resource has been modified is not important, only that the resource has been revised. Peter Graham, for example, asks, ‘how can a reader be sure that the document being used is the one intended?’ [38]. In a similar vein, Luciana Duranti – in an archives context – has defined ‘authenticity’ as proving that a document is what it claims to be [39]. In a networked information environment, ‘authentication’ is, therefore, concerned with both establishing identity and ensuring data integrity.

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 55, No. 1, January 1999

It is possible that the adoption of unique identifiers will help solve some of these problems, particularly those related to the version control of digital resources. In addition, the rationale that underlies the development of the DOI is also related to authenticating identity in that the use of unique identifiers and an associated resolution service would enable publishers, or other agencies, to manage access and copyright functions.

The project was only interested in a subset of these issues, primarily in a bibliographic context. BIBLINK adopted a pragmatic working definition of 'authentication': a guarantee that a piece of metadata actually describes a given electronic publication, and only that publication. There needed to be an authenticated one-to-one relationship between an electronic publication and its metadata [40]. The BIBLINK study concluded that any authentication mechanism would have to work in accordance with two specific models:

- Publications that are stored in a controlled environment, e.g. off-line publications like CD-ROMs that are deposited in a national library. Mechanisms are needed to link the metadata given at the time of creation with the item it describes. Any change in the original item, e.g. migration to another format for preservation, should be noted in the metadata.
- Publications that are stored in an uncontrolled environment, e.g. online resources like Web pages that would be managed outside the national library context. Metadata for these resources would be created and authentication mechanisms would have to be devised to ensure that this metadata actually matches with the distributed resources themselves.

The BIBLINK study of authentication included a review of projects and technologies, and this looked at authentication methods and techniques used in a variety of electronic document delivery and copyright management projects and metadata initiatives. The study concluded that version control was important. While traditional publishers had developed elaborate practices for identifying reprints, revisions and new editions of print publications, there were no equivalent standards for electronic publications. In the absence of any agreed standards in this area, the BIBLINK study recommended the use of hashing techniques to meet the project's requirements with regard to version control and document integrity. Hashing is a cryptographic technique for checking the integrity of data by the production of a hash value or checksum. A checksum has been described as a 'fixed length block functionally dependent on every bit' of a resource, so that different resources would have different checksums, 'with high probability' [41, p. 147]. An authentication checksum would be computed from a resource and would then be added to the descriptive metadata itself. When a user retrieves this resource at a later date, this checksum could be computed again and then compared with the checksum recorded in the metadata. If the two checksums agree, there can be confidence that the metadata does refer to that version of a resource and none other. On the other hand, if the resource has been updated or manipulated in some other way, the user will be aware of this, even though it will not provide information on the precise nature of these changes.

BIBLINK could only deal with the practical problems of authentication as they relate to the project itself. Identifying other metadata that could be used to

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 55, No. 1, January 1999

authenticate digital resources will be an important part of other projects and implementations that deal with the wider issues of digital preservation. Metadata that can help with version control and mechanisms that can be used to check the integrity of resources may become important parts of bibliographic standards in the future.

THE DEMONSTRATOR MODEL

The BIBLINK demonstrator consists of a virtual workspace that will act as a working environment and as a database (Figure 4). Publishers will first create metadata that can be transferred to the workspace. Once there, publishers and other participants will be able to retrieve, revise and delete records. The BIBLINK workspace will additionally perform all relevant format conversions. In the first instance the publishers' metadata will be converted into a UNIMARC record, stored, and then this will be converted into a national MARC format. The record in both its original format and in the national MARC format would then be forwarded to the relevant national bibliographic service where it can be added to its local MARC-based database and enhanced. The enhanced national MARC record can then be sent back to the BIBLINK workspace where it will be stored, and then converted into an enhanced UNIMARC record. An enhanced BIBLINK Core record can then be created from this and a copy sent to the publisher.

CONCLUSIONS

BIBLINK has supported a structured investigation by a number of European national libraries into the bibliographic control of electronic publications. The role of national bibliographies is evolving and BIBLINK has highlighted some of the issues. The project has considered which publications constitute a nation's intellectual record, what can be considered as a publication in the Internet environment, and who are the publishers. It has examined requirements

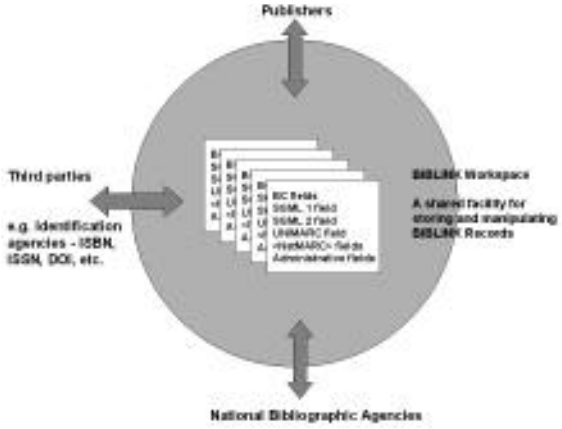

Figure 4. *BIBLINK workspace model*

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 55, No. 1, January 1999

for metadata relating to electronic resources compared to traditional national bibliographic records.

In future each national library will need to consider which publications it wants to record. Inevitably this will be affected by considerations of national heritage and publishing practice. Some national libraries may see themselves as having a role as an authorising agency for assignment of unique identifiers, others may leave this task to other agencies.

Generating links, electronic and otherwise, with publishers and other agents in the bibliographic information chain is becoming more important in the digital age. These links will impact the future development of standards for bibliographic information, standards such as ISBD and formats such as UNIMARC increasingly will be influenced by the requirements of wider interests including those of publishers, web site managers and other user communities. At the very least improved interoperability with simple structured metadata formats like Dublin Core will become desirable.

Bibliographic control of electronic publications is an ambitious task, unlikely to be achieved by national libraries working in isolation. There will be varying degrees of co-operation with other services to achieve some level of bibliographic control. These agencies may be from the commercial, education and public sectors. This already happens to a certain extent with printed publications, e.g. in the UK there are both commercial bibliographic agencies and the copyright academic libraries contributing to the process. For electronic resources one might envisage a role for specialist subject services, commercial search services, as well as from publishers themselves. The National Library of Australia's position paper on access to electronic publications states:

> Whereas libraries have built up close working relationships with traditional print publishers involving a mutual appreciation of the respective roles of print publishing and libraries, it is likely that electronic publishing will involve a new range of players, and that new relationships and understandings will need to be forged. Libraries will need to publicise their roles and interests in electronic publishing, and listen to the needs and concerns of producers. Areas of joint interest and activity need to be identified [3].

BIBLINK has focused on pragmatic solutions to enable a demonstrator to be established in the time-scale of the project. We hope that lessons learnt within the project may contribute to the future definition of roles and agreement on best practice.

ACKNOWLEDGEMENTS

The work described in this paper has been supported by the BIBLINK project, funded under the EU Telematics Application Programme. UKOLN is funded by the Joint Information Systems Committee (JISC) of the UK Higher Education Funding Councils, the British Library Research and Innovation Centre and by project funding from several sources. The authors remain responsible for any views expressed in this article.

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 55, No. 1, January 1999

REFERENCES

1. Line, M.B. Back to basics for national libraries? *Alexandria*, 7(1), 1995, 1–2.
2. Hoare, P. Legal deposit of electronic publications and other non-print material: an international overview. *Alexandria*, 9(1), 1997, 59–79.
3. National Library of Australia. *National strategy for provision of access to Australian electronic publications: a National Library of Australia position paper.* Canberra: National Library of Australia, December 1996. http://www.nla.gov.au/policy/paep.html (visited 30 July 1998).
4. Plassard, M.-F. The IFLA Core Programme for Universal Bibliographic Control and International MARC (UBCIM): recent developments and current state. *Alexandria*, 6(2), 1994, 145–153.
5. Anderson, D. An international framework for national bibliographic development: achievement and change. *Library Resources and Technical Services*, 30(1), 1986, 13–22.
6. Line, M.B. National self-sufficiency in an electronic age. In: Helal, A. and Weiss, J., eds. *Electronic documents and information: from preservation to access. 18th International Essen Symposium.* Essen: Universitätsbibliothek Essen, 1996, 170–192.
7. Mackenzie Owen, J.S. and Walle, J. van de. *Deposit collections of electronic publications.* Luxembourg: European Commission, DG XIII-E/4, 1996.
8. Lehmann, K.-D. Making the transitory permanent: the intellectual heritage in a digitized world of knowledge. *Daedalus*, 125(4), 1996, 307–329.
9. Bourne, R. Bibliographic standards. In: Bourne, R., ed. *Serials librarianship.* London: Library Association, 1980, 187–197.
10. Davis-Brown, B. and Williamson, D. Cataloging at the Library of Congress in the digital age. *Cataloging & Classification Quarterly*, 22(3/4), 1996, 171–196.
11. Williamson, D. Text Capture and Electronic Conversion. Presentation given at the Library of Congress, October 13, 1994. In: *Proceedings of the Seminar on Cataloging Digital Documents, October 12–14, 1994, University of Virginia Library, Charlottesville and the Library of Congress*. http://lcweb.loc.gov/catdir/semdigdocs/david.html (visited 30 July 1998).
12. Morris, S. Electronic cataloging speeds procedure. *A periodic report from The National Digital Library Program, The Library of Congress*, 6, March 1996. http://lcweb.loc.gov/ndl/march-96.html#cataloging (visited 30 July 1998).
13. *BIBLINK: linking publishers and national bibliographic services.* 1996. http://hosted.ukoln.ac.uk/biblink/ (visited 30 July 1998).
14. Zillhardt, S. CoBRA: une action concertée entre bibliothèques nationales. *Bulletin des bibliothèques de France*, 41(1), 1996, 66–69.
15. Lehmann, K.-D. European national libraries and the CoBRA Forum of the EU Libraries. *Alexandria*, 8(3), 1996, 155–166.
16. Curwen, A.G. UNIMARC and international record exchange: an overview of recent projects and developments. *Program*, 31(3), 1997, 227–238.
17. Iljon, A. The European Libraries Programme: an overview. *Program*, 29(4), 1995, 361–377.
18. Heery, R., et al. *Study of metadata.* BIBLINK D1.1, December 1996. http://hosted.ukoln.ac.uk/biblink/wp1/d1.1/ (visited 30 July 1998).

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 55, No. 1, January 1999

19. Dempsey, L. and Heery, R. *A review of metadata: a survey of current resource description formats.* DESIRE D3.2, March 1997. http://www.ukoln.ac.uk/metadata/desire/overview/ (visited 30 July 1998).
20. Dempsey, L. *Bibliographic records: use of data elements in the book world.* Bath: Bath University Library, Centre for Bibliographic Management, 1989. (BNBRF Report 40)
21. Dempsey, L. Users' requirements of bibliographic records: publishers, booksellers, librarians. *Aslib Proceedings*, 42(2), 1990, 61–69.
22. Curwen, A.G. International Standard Bibliographic Description. In: McIlwaine, I.C., ed. *Standards for the international exchange of bibliographic information: papers presented at a course held at the School of Library, Archive and Information Studies, University College London, 3–18 August 1990.* London: Library Association, 1991, 73–81.
23. Gredley, E. and Hopkinson, A. *Exchanging bibliographic data: MARC and other international formats.* London: Library Association, 1990.
24. ISBD(CF) Review Group. *ISBD(ER): International Standard Bibliographic Description for Electronic Resources.* Munich: K.G. Saur, 1997. (UBCIM Publications, n.s., 17)
25. Dempsey, L. and Heery, R. Metadata: a current view of practice and issues. *Journal of Documentation*, 54(2), 1998, 145–172.
26. Pira International. *Simplified SGML for Serial Headers (SSSH).* Pira, 1996.
27. Heery, R., et al. *Format conversion feasibility.* BIBLINK D4.1, September 1997. http://hosted.ukoln.ac.uk/biblink/wp4/d4.1/ (visited 30 July 1998).
28. Olson, N.B., ed. *Cataloguing Internet resources: a manual and practical guide.* 2nd ed. Dublin, Ohio: Online Computer Library Center, 1997. Also available at: http://www.purl.org/oclc/cataloging-internet (visited 30 July 1998).
29. Paskin, N. Information identifiers. *Learned Publishing*, 10(2), 1997, 135–156.
30. Høgås, H., Werf, T. van der and Powell, A. *Identification.* BIBLINK D2.1, May 1997. http://hosted.ukoln.ac.uk/biblink/wp2/d2.1/ (visited 30 July 1998).
31. Sollins, K. and Masinter, L. *Functional Requirements for Uniform Resource Names.* RFC 1737, December 1994. http://info.internet.isi.edu:80/in-notes/rfc/files/rfc1737.txt (visited 30 July 1998).
32. Lynch, C., Preston, C. and Daniel, R. *Using existing bibliographic identifiers as Uniform Resource Names.* RFC 2288, February 1998. http://info.internet.isi.edu:80/in-notes/rfc/files/rfc2288.txt (visited 30 July 1998).
33. Sun, S.X. *Handle System: a persistent global naming service: overview and syntax.* Internet-Draft, 16 July 1998. http://www.ietf.org/internet-drafts/draft-sun-handle-system-01.txt (visited 30 July 1998).
34. Lynch, C. Identifiers and their role in networked information applications. *ARL: A Bimonthly Newsletter of Research Library Issues and Actions*, 194, October 1997. http://www.arl.org/newsltr/194/identifier.html (visited 30 July 1998).
35. Bide, M. *In search of the Unicorn: the Digital Object Identifier from a user perspective*, rev. ed. London: Book Industry Communication, 1998. Also available at: http://www.bic.org.uk/bic/unicorn2.pdf (visited 30 July 1998). (BNBRF Report 89)

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 55, No. 1, January 1999

36. Lynch, C.A. Authentication and authorization. Part 1: the changing role in a networked information environment. *Library Hi Tech*, 15(1–2), 1997, 30–38.
37. Brassard, G. *Modern cryptology: a tutorial.* Berlin: Springer-Verlag, 1988. (Lecture notes in computer science, 325)
38. Graham, P.S. *Intellectual preservation: electronic preservation of the third kind.* Washington, DC: Commission on Preservation and Access, 1994. Also available at: http://www.clir.org/pubs/reports/graham/intpres.html (visited 30 July 1998).
39. Duranti, L. Reliability and authenticity: the concepts and their implications. *Archivaria*, 39, 1995, 5–10.
40. Werf, T. van der, et al. *Authentication.* BIBLINK D6.1, September 1997. http://hosted.ukoln.ac.uk/biblink/wp6/d6.1/ (visited 30 July 1998).
41. Denning, D.E.R. *Cryptography and data security.* Reading, Mass.: Addison-Wesley, 1982.

(*Revised version received 4 August 1998*)

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib