4
Key Issues in Digital Library Interoperability

Renato Iannella, Nigel Ward
Research Data Network Cooperative Research Centre
DSTC Pty Ltd, Level 7, Gehrmann Laboratories
The University of Queensland, QLD 4072, AUSTRALIA
Email: {renato,nigel}@dstc.edu.au

Kaveri Chakrabarty
National Library of Australia, Canberra, ACT 2600, AUSTRALIA
Email: k.chakrabarty@nla.gov.au

Chris Curtis
CSC Australia, Suite 5, Trevor Pearcey House
Traeger Court, Fern Hill Park, Bruce, ACT 2617, AUSTRALIA
Email: c.curtis@nla.gov.au

Abstract

To cope with the explosion of electronically available information, Digital Libraries must interact if they are to meet the information requirements of their clients. This paper gives an overview of some of the key issues that digital projects will need to address so as to reach an acceptable level of interoperability.

Keywords

Digital Libraries, Interoperability, Metadata, Scalability, Z39.50, Resource Discovery, WWW.

1 INTRODUCTION

Many authors have suggested that libraries can no longer just be physical spaces which allow browsing and borrowing of materials, they must also provide access to the exploding range of electronically available information. In response to this realisation, there are many Digital Library (DL) projects, worldwide, that are attempting to establish digital collections, or at least digital access to the catalogues of collections. Interactions between these DLs is necessary to make the problem of finding relevant digital information tractable.

This paper gives an overview of some of the key issues that digital projects will need to address so as to reach an acceptable level of interoperability. These issues include; User Interfaces, Standards for Sharing Information, Metadata, and Discovery of Digital Libraries. By addressing these issues, DLs will be able to seamlessly search other catalogues, retrieve documents, and use the services of other libraries. An example DL, the National Document & Information Service (NDIS) Project, a joint initiative of the National Libraries of Australia and New Zealand, the Open Information Locator (OIL) research project from the DSTC, and National Library of Australia research into user interfaces are discussed in the light of these issues.

N. Terashima et al. (eds.), *Advanced IT Tools*

Digital Library Issues

One of the main factors that delineate a DL from a large distributed repository of online data such as that found in the World-Wide Web (WWW) environment at present, is that a DL embodies both information and services. Current libraries are charged with the selection, collection, organisation, access, and preservation of physical materials that are relevant to the libraries' responsibilities. DLs have the same responsibilities of their paper library equivalents but must manage and organise electronic collections and catalogues. Also, they must cater for the needs of remote users who access the DL electronically, without the ability to interact at the human level. In many cases, DLs will themselves access external sources, which they have little control over, that will be integrated into their own collections. DLs must also address open problems in the electronic area including management of intellectual property and administration of charging regimes in areas of different services.

A DL also needs to provide effective retrieval mechanisms to the entire range of its clients, offering a smooth transition of methods to suit the level of expertise and the nature of requirements of the clients. DLs are expected to hold both collections as well as catalogues and indices of collections, paper and electronic, for the long term future, and to present to the user an integrated coherent view of the total collection of information.

There are many projects that are developing large-scale DLs. Of particular note is the US Digital Library Initiative which has sponsored six projects. Each project has taken a different approach or focused on particular areas of DL research. See Schatz & Chen (1996) for a detailed analysis of the projects. The Information Infrastructure Technology and Applications Working Group has also identified key areas that require significant research and development investigations (Lynch & Garcia-Molina, 1995). These include; interoperability, description of objects and repositories, collection management and organisation, user interfaces and Human-Computer Interaction, and economic, social, and legal issues.

2 USER INTERFACE

Since DLs will be providing access to a large database of information, some innovative metaphors need to be developed to support effective browsing and searching. Ideally the details of the DL being accessed should be hidden from the user. For this reason, consistent user interfaces across DLs should be supported.

Within the community of people who will use DLs there is a wide range of experience and expectation. Some users will be professional librarians, researchers or search intermediaries who often demand sophisticated and powerful interfaces that allow them to precisely specify their information needs and give maximum precision and recall (ie the system should return *only* documents that satisfy the query and should return *all* the documents that satisfy the query). Other less experienced users require the user interface of a DL be simple to use and tolerant of imprecise queries. The user interface of a DL must be flexible enough to cater for all these needs.

It is becoming common for DL interfaces to provide access to a number of databases, which are often widely dispersed geographically, run on different systems, have different search syntaxes, format search results differently, have different charging schemes, and vary in the precision of query they allow. It is the role of the DL interface to resolve all these differences and present a common view of all the services to the user. Ideally, the location and nature of the databases searched should be transparent to the novice user; the DL interface should search whichever databases are most appropriate for satisfying the users query without the user needing to know that separate databases are involved.

The NDIS User Interface

NDIS provides both a custom developed graphical user interface and a WWW interface. The WWW interface is seen as a long term strategic direction for access to NDIS, while in the shorter term the custom interface will provide functionality that is not yet available or easy to implement using WWW, such as cataloguing functions and multiple window support (see Figure 1). Emerging WWW technology, such as the frames features of Netscape 2.0 and downloadable

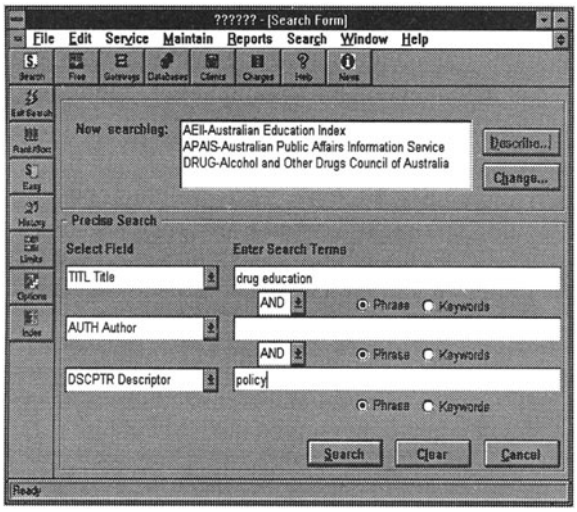

Figure 1: NDIS Client Interface

applets will mean that all NDIS functions will be able to be provided through WWW browsers in the not too distant future.

The NDIS system is providing three different interfaces to the systems search facilities: *Easy*, *Precise* and *Command*. The Easy interface allows the user to select which databases to search from the range of databases available then has a free text entry box in which the user can specify the search terms. If no databases are selected, the system defaults to searching a system defined subset of the whole information resource. The system uses a number of strategies to find documents based on the terms entered by the user. Initially they are treated as a phrase. If this is unsuccessful, the system reformats the query using the AND operator between each word. If this is also unsuccessful, OR is used, followed by a range of increasingly fuzzy matching techniques. This process has been designed to give the maximum chance of success to any query entered through the easy interface.

The Precise interface is a graphical interface that allows the user to select which fields to search, to specify Boolean operators to combine search terms and to specify a range of proximity and truncation requirements. This allows the experienced user to precisely specify their search and the system executes the search as specified by the user.

The Command interface also allows users to precisely specify their search using the Common Command Language (NISO standard Z39.58), rather than using the point and click Precise

graphical interface. This is considered a much faster way for very experienced information professionals to specify their search requirements.

The NLA Interface

Concurrent with NDIS, the NLA have been developing an interface to transparently access multiple databases of publicly available information, with special emphasis on information in the Australian Internet domains. This interface forms part of the provision for information access to NLA's clients and complements the chargeable access services through NDIS.

This interface has been implemented to invoke multiple search engines in parallel and to merge multiple result sets into a single set to the user, with an optional document summarising facility. Like NDIS, this interface accepts a single search syntax and returns a common view from all the services to the user, and is pitched at a level between Easy and Precise NDIS options. The current implementation uses a number of Australian indexing services and the US based WWW indexing service Alta-Vista, and incorporates some multimedia databases such as the NLA's image collection. This interface has been developed for both for NLA's own DL collections and as an entry point to a Whole-of-Government Australian electronic collection.

The OIL User Interface

The OIL project has been investigating new interface paradigms suitable for dealing with the large volume of information contained in Digital Libraries. OIL uses a HyperIndex Browser (HIB) as its user interface technology, which provides a *browsing* interface to large information spaces. In particular, the HIB has been implemented to browse large result sets from WWW index services (such as Lycos and Alta-Vista). The HIB works by processing the results from a user's queries, extracting relevant keywords, and presenting this list to the user. The user then has the opportunity to refine or enlarge their query. This is repeated until the user is satisfied with the current focus of the search and can then obtain the actual results of the query.

Figure 2 shows an example of the HIB interface. In this case, the query for 'Digital Libraries' is shown with a list of possible refinements and enlargements. For further information on HIB within the OIL project see (Iannella *et al*, 1995).

3 STANDARDS FOR SHARING INFORMATION

The key information sharing standard for DLs is Z39.50, a standard information retrieval protocol that has been developed with input from a range of international information providers and users (ANSI, 1995). Z39.50 provides a flexible framework for an information client to request information retrieval tasks on databases accessible through an information server. The library community is using Z39.50 to provide remote access to both catalogues and digital collections. In the broader arena of public sector information, Z39.50 has been adopted by the U.S. Government Information Locator Service (GILS, 1994) initiative, and is being considered for use by the Australian Office of Government Information Technology initiatives.

Since Z39.50 is a standard protocol, it does solve some library interoperability issues and its scope is expanding with the release of newer versions, however, it does not yet guarantee interoperability. Due to its flexibility, different information providers can provide the same information in different ways using Z39.50. For this reason, standard ways of sharing information using Z39.50 are being developed.

Clearly, DLs need to cooperate to make available their catalogue of information to the rest of the world. They need to actively support international agreements on the content of such resources. The Catalogue Interoperability Protocol (CEOS, 1995) and the Author-Title-Subject

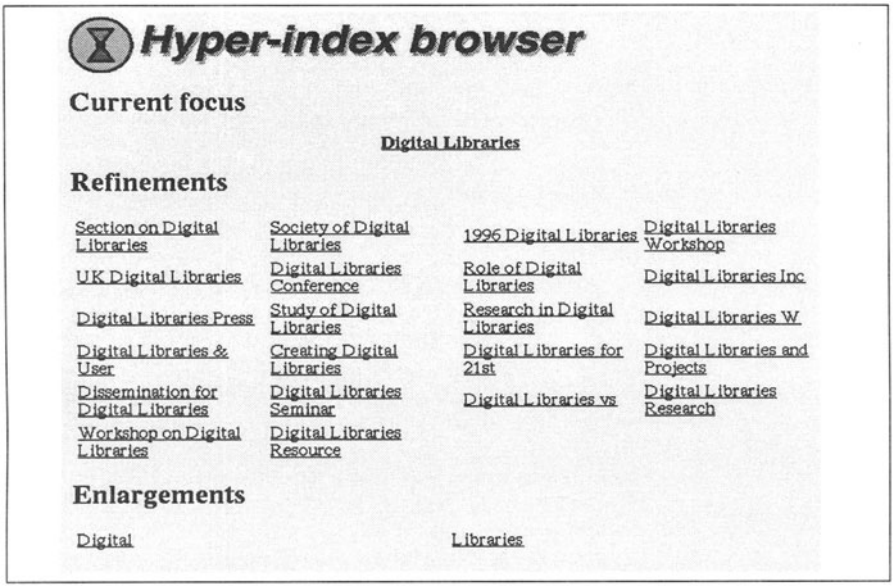

Figure 2: HyperIndex Browser Interface

Profile (OIW/SIGLA, 1995) are agreements which provide guidelines for using Z39.50 to provide access to library (and other) catalogues. By complying with these guidelines DLs guarantee a level of understanding of the structure of their catalogues.

Some DLs provide access to digitised collections as well as digitised cataloguing information. These collections often contain organisational structure and descriptive information which aids user navigation. The Z39.50 Digital Collections Profile (Library of Congress, 1996) is a standard for allowing this navigation information to be shared between DLs. By complying with this profile a DL allows its digital objects to be accessed and browsed in a standard way.

Standards in the NDIS Project

NDIS will be providing Z39.50 client and server support. This will enable users of NDIS to search other DL collections using the NDIS interface. Users of other DLs will be able to access data held in NDIS using their own interface. The NDIS system will initially contain over 40 differently structured databases. The system will allow transparent searching across all these databases. In a future release, NDIS will use the Z39.50 protocol to allow transparent searching of external databases such as the Library of Congress, OCLC and the British Library.

4 METADATA

To support interoperability, DLs need to adequately describe their contents. This description, or metadata, is crucial to support scalable discovery of information. The metadata needs to effectively and concisely describe the contents of the library. This will enable a client to decide if a particular DL contains information related to a user's information need.

There are two levels of metadata; semantic and technical. Semantic metadata describes the intellectual content of the object and attempts to provide the reason or meaning of the resource. Technical metadata describes physical or electronic characteristics about the object and usually describes how to access the object and what formats it is available in. Metadata, in relation to DLs, also can be classified into two types; metadata that describes document-like objects (DLOs) and metadata that describes *collections* of DLOs.

Metadata for document-like objects

For a considerable time now, librarians have used well evolved cataloguing structures to describe their holdings (AACR, 1988) that allow for very precise searching and identification of items. The process of creating metadata for a publication is labour intensive and requires much intellectual input by trained cataloguers. A number of approaches are being adopted to reduce the cost of metadata creation including the growing move towards *shelf-ready* books. These are packages that contain a book plus an electronic metadata record for the book that is prepared by the publisher. Thus, the cost of the creation of the metadata can be shared among many purchasers of the book.

There are also moves towards automatic metadata creation. With a growing number of documents being published electronically using SGML, it is possible to automatically identify metadata elements such as titles and authors. As more and more documents are stored electronically and the full text of the document can be indexed and searched, there is increased need for concise metadata to define the contents of documents - so that index sizes can be contained and retrieval focus improved.

There are a number of metadata standards relating to DLOs. One of the emerging standards is the Dublin Core metadata element set (Weibel *et al*, 1995). The Dublin Core elements primarily describe the semantic aspects of DLOs. Current work on Dublin Core include the standardisation of a syntax protocol (Burnurd *et al*, 1996) and a transport infrastructure for metadata packages. Most of this work is the result of the Warwick Dublin Core Workshop (Lagoze, 1996).

The BIB-1 attribute set in the Z39.50 standard is another key metadata standard for cataloguing DLOs. It has been designed to implement searching using metadata developed by librarians.

Document metadata in the NDIS project

NDIS holds metadata separately from the physical structure of the database. As metadata standards continue to evolve, NDIS will be able to adapt to changing requirements without program or database changes. Quality indexing is also a crucial element in locating relevant information. NDIS and other National Library projects are deploying a text retrieval engine to index both on metadata and on full texts

In the NDIS system metadata from each (full text) database is being mapped into a common set of metadata attributes, a subset of Dublin Core. This allows users to query across differing databases using a single view of the data without requiring the users to know the structure of each database and without requiring the database producers to conform to a single standard.

Collection metadata

Another part of a DLs metadata is a description of the technical aspects on how to access and interact with the services offered by that library. This would include information on how to query the library and what the format of the results of the search will be. Other information may include, the costs of the searches and the operating hours of the service. A key aspect of this interaction is that it occurs dynamically, and can be negotiated at run-time. The Explain facility

of the Z39.50 protocol (1995 version) is a rudimentary system that can support this type of interoperability.

Collection metadata in the OIL project

The OIL project has developed some standard metadata descriptions for WWW index services (such as Lycos and Alta-Vista). The metadata is used to dynamically access these services in the OIL prototype. The current metadata is focused on the technical aspects of the services. The next stage is to investigate the semantic metadata used to describe these services. Ideally, the metadata would be maintained by the service providers, but in the interim, the DSTC is supporting this to enable the OIL prototype to experiment with various metadata strategies.

5 DISCOVERY

To participate in a (possibly) global collection of DLs, individual DLs need to be discovered. The yellow-pages style lookup provided by Resource Discovery systems are one means of finding such libraries. Effective and *scalable* advertising mechanisms, naming of DLs, and dynamic discovery of DLs, must be investigated.

OIL Project approach

The research focus of OIL is on the scalability issues associated with finding and accessing a large and growing number of DLs. Similar scalability issues are being researched by the open distributed processing (ODP) community. The DSTC is investigating the ODP Trader (ISO 1994) and interworking technologies as a solution to this problem, and implementing a resource discovery Trader based on the X.500 directory service. DLs providers register details of their service as a service offer within such a Trader. Federation of these interworking Traders provides a solution to scalability.

The federated network of resource discovery Traders is fundamental to a solution for the problem of query routing. The Trader will be used to determine what DL sources to query by returning service offers from information sources that are relevant to a query. This solution also has metadata implications as it is necessary to provide a characterisation of the information providers.

6 SUMMARY

DLs will be an important component for future global information provision. They will play a significant role in the provision and access to information for the public. The DL developers must seize this opportunity and exploit current and future technologies to provide one seamless interface to their resources. To be successful, the issues described in this paper must be faced and addressed by the DL community.

7 ACKNOWLEDGEMENTS

The work reported in this paper has been funded in part by the Cooperative Research Centres Program through the Department of the Prime Minister and Cabinet of Australia. The authors acknowledge the work of John Evershed, of Project Computing Pty Ltd, on the NLA Interface.

8 REFERENCES

AACR (1988) Anglo-American Cataloguing Rules, Second Edition - 1988 Revision, Edited by Michael Gorman and Paul W. Winkler, American Library Association, 1988

ANSI (1995) ANSI/NISO Z39.50-1995 (Versions 2 and 3) Information Retrieval: Application Service Definition and Protocol Specification, 1995.

Burnard, L & Miller, E & Quin, L & Sperberg-McQueen, CM. (1996) A Syntax for Dublin Core Metadata - Recommendations from the Second Metadata Workshop, 1996. *<url: http://info.ox.ac.uk/~lou/wip/metadata.syntax.html>*

CEOS (Committee on Earth Observation Satellites) Working Group on Data. (1995) Catalogue Interoperability Protocol (CIP) Specification. *<ftp://styx.esrin.esa.it/pub/od/CIP/>*.

GILS (1994) The Government Information Locator Service. May 2, 1994. *<url: http://www.usgs.gov/gils/gilstoc.html>*

Iannella, Renato & Ward, Nigel & Wood, Andrew & Sue, Hoylen & Bruza, Peter. (1995) Digital Libraries and the Open Information Locator Project. DSTC Technical Report *<http://www.dstc.edu.au/RDU/reports/OIL/oil-dl.ps>*

ISO Trader (1994), Rec.X.9tr/Draft ODP Trading Function ISO/IEC 13235: 1994/Draft ODP Trading Function. ISO/IEC JTC1/SC21, N-9122, July, 1994.

Lagoze, Carl. (1996) The Warwick Framework - A Container Architecture for Aggregating Metadata Objects, 1996 *<url: http://cs-tr.cs.cornell.edu/~lagoze/warwick.htm>*

Library of Congress (1996) Z39.50 Profile for Access to Digital Collections. *<url: http://lcweb.loc.gov/z3950/agency/profiles/collections.html>*

Lynch, Clifford & Garcia-Molina, Hector (1995). Interoperability, Scaling, and the Digital Libraries Research Agenda. IITA Digital Libraries Workshop, August 22, 1995. *<http://www-diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html>*

OIW/SIGLA (Open Implementors Workshop, Special Interest Group on Library Automation) (1995). ATS-1 (Author, Title, Subject) Profile. *<http://lcweb.loc.gov/z3950/agency/profiles/ats.html>*.

Schatz, Bruce & Chen, Hsinchun. (1996) Building Large-Scale Digital Libraries. *IEEE Computer*, May 1996.

Weibel, S. & Godby, J & Miller, E & Daniel, R. (1995) OCLC/NCSA Metadata Workshop Report, Dublin, Ohio, USA, March 1-3, 1995. *<url: http://www. oclc.org:5047/oclc/research/conferences/metadata/dublin_ core_report.html>*

9 BIOGRAPHY

Renato Iannella is a Senior Research Scientist at the Distributed Systems Technology Centre (DSTC) based in Brisbane, Australia. At the DSTC, he leads the Resource Discovery Unit which investigates technologies used in the discovery and retrieval of Internet resources on globally distributed networks.

Nigel Ward is a Senior Research Scientist at the Distributed Systems Technology Centre (DSTC) based in Brisbane, Australia. Nigel's work within the Resource Discovery Unit has focused on Internet Information Retrieval Protocols, particularly the Z39.50 information retrieval protocol.

Kaveri Chakrabarty is a Senior Information Technology Officer at the National Library of Australia (NLA). At the NLA, she is responsible for the Host Systems unit that supports hardware, Operating Software and database functions on MVS and Unix platforms, and develops and integrates technical options for new IT services. Her current roles include leading an Internet Services group and providing technical advice to the NDIS project on System Architecture and platform support.

Chris Curtis works as a consultant for CSC Australia. His current position is System Architect for the NDIS project for the National Libraries of Australia and New Zealand. In this role he is responsible for the overall design and integration of all aspects of the system. He has worked in the computer industry since 1977, mainly in the areas of application development, database design, and information modelling.