DESIRE: Project Deliverable

Project Number:	RE 1004 (RE)	
Project Title:	DESIRE - Development of a European Service for Information on Research and Education	
Deliverable Number:	D3.2	
Deliverable Title:	Specification for resource description methods. Part 1. A review of metadata: a survey of current resource description formats.	
Version Number	1.0 (19 March 1997)	
Deliverable Type:	PU	
Deliverable Kind:	RE	
Principal Author:	Name	**Lorcan Dempsey and Rachel Heery**
	Address	UKOLN, University of Bath, Bath, BA2 7AY, UK
	E-Mail	L.Dempsey@ukoln.ac.uk, R.M.Heery@ukoln.ac.uk
	Telephone	+44 (0)1225 826580
	Fax	
Other Authors:	**Martin Hamilton, Debra Hiom, Jon Knight, Traugott Koch, Marianne Peereboom and Andy Powell**	
Deliverable URL(s):		
Abstract:	This study provides background information to the DESIRE project to enable the implications of using particular metadata formats to be assessed. Part I is a brief introductory review of issues including consideration of the environment of use and the characteristics of metadata formats. A braod typology of metadata is introduced to provide a framework for analysis. Part II consists of an outline of resource description formats in directory style. This includes generic formats, but also, to give an indication of the range of development, domain-specific formats. The focus is on metadata for 'information resources' broadly understood rather than on the variety of other approaches which exist within particular scientific, engineering and other areas.	
Keywords:	metadata, BibTex, CIMI, Dublin Core, EAD, EEVL, EELS, FGDC, GILS, IAFA, whois++, ICPSR, LDAP, MARC, USMARC, UKMARC, UNIMARC, PICA+, SOIF, TEI,URC, Warwick Framework	

PART II: Executive Summary

This document has been prepared as a working paper in the Indexing and Cataloguing work package of the DESIRE project (WP3), funded by the European Union as part of the Telematics for Research area of the Fourth Framework Programme. A companion document *Resource description: initial recommendations for metadata formats*

<URL:http://www.ukoln.ac.uk/metadata/DESIRE/recommendations>

lays out recommendations for metadata use within the subject-based information gateways within DESIRE.

The DESIRE project will use a generic metadata format for the records in the subject-based information gateways. There are a number of options for this format. This study provides background information which allows the implications of using particular formats to be assessed. Part I is a brief introductory review of issues. Part II provides an outline of resource description formats in directory style. This second part includes generic formats, but also, to give an indication of the range of development, domain-specific formats. The intention is not to be comprehensive, but to give sufficient examples to support understanding of a rapidly developing environment. The focus is on metadata for 'information resources' broadly understood; a variety of other approaches exist within particular scientific, engineering and other areas.

Metadata is data which describes attributes of a resource. A more formal definition is:

> metadata is data associated with objects which relieves their potential users of having to have full advance knowledge of their existence or characteristics.

It is recognised that in an indefinitely large resource space, effective management of networked information will increasingly rely on effective management of metadata. The need for metadata services is already clear in the current Internet environment. As the Internet develops into a mixed information economy deploying multiple application protocols and formats this need will be all the greater. Metadata is not only key to discovery, it will also be fundamental to effective use of found resources (by establishing the technical or business frameworks in which they can be used) and to interoperability across protocol domains.

Thus, the nature of the problem to be solved suggests a variety of solutions. It is unlikely that some monolithic metadata format will be universally used. This is for a number of more or less well known reasons, not least the investment represented by legacy systems in terms of technology and human effort.

In Part I of this report we examine some characteristics of the environment in which network information of interest to European researchers is being created and some of the factors which are influencing the development of metadata services.

Part II of this report describes a range of metadata formats under a number of standard headings.

PART III

Contents

PREFACE

This document has been prepared as a working paper in the Indexing and Cataloguing work package of the DESIRE project (WP4), funded by the European Union as part of the Telematics for Research area of the Fourth Framework Programme.

This particular task is co-ordinated by UKOLN and the involved partners are UKOLN at the University of Bath, Loughborough University, Koninklijke Bibliotheek, National Library of the Netherlands, and SOSIG at the University of Bristol.

In preparing the material, the authors contacted a range of standards developers and implementors for further information or clarification. We would like to thank them for their kind co-operation and assistance. Any errors or misinterpretations remain the authors'. Those contacted included:

- CIMI (Bill Moen)
- CDWA (Murtha Baca)
- EAD (Daniel Pitti)
- EELS (Traugott Koch)
- EEVL (Malcolm Moffat)
- FGDC (Doug Nebert)
- ICPSR (Ann Green, Mary Vardigan)
- UNIMARC (Brian Holt)

The authors acknowledge the helpful comments of the peer reviewers in the preparation of the final report.

A companion document *Resource description: initial recommendations for metadata formats* <URL:http://www.ukoln.ac.uk/metadata/DESIRE/recommendations> lays out recommendations for metadata use within the subject-based information gateways within DESIRE.

PART I - AN OVERVIEW OF RESOURCE DESCRIPTION ISSUES

SCOPE

The DESIRE project will use a generic metadata format for the records in the subject-based information gateways. There are a number of options for this format. This study provides background information which allows the implications of using particular formats to be assessed. Part I is a brief introductory review of issues. Part II provides an outline of resource description formats in directory style. This includes generic formats, but also, to give an indication of the range of development, domain-specific formats. The intention is not to be comprehensive, but to give sufficient examples to support understanding of a rapidly developing environment. The focus is on metadata for 'information resources' broadly understood; a variety of other approaches exist within particular scientific, engineering and other areas.

METADATA AND ITS USES

Metadata is data which describes attributes of a resource. Typically, it supports a number of functions: location, discovery, documentation, evaluation, selection and others. These activities may be carried out by human end-users or their (human or automated) agents.

A more formal definition is:

> metadata is data associated with objects which relieves their potential users of having to have full advance knowledge of their existence or characteristics.

It is recognised that in an indefinitely large resource space, effective management of networked information will increasingly rely on effective management of metadata. The need for metadata services is already clear in the current Internet environment. As the Internet develops into a mixed information economy deploying multiple application protocols and formats this need will be all the greater. Metadata is not only key to discovery, it will also be fundamental to effective use of found resources (by establishing the technical or business frameworks in which they can be used) and to interoperability across protocol domains.

Part II of this report describes a range of metadata formats. It is unlikely that some monolithic metadata format will be universally used. This is for a number of more or less well known reasons, not least the investment represented by legacy systems in terms of technology and human effort. In addition the variety of record formats represent an attempt to meet the diverse requirements of the different communities. The various communities involved in resource description have vested significant effort in developing specialised structures to enable rich record descriptions to be created to fulfil the requirements of their particular domain. These structures are embodied in systems. In addition the people who maintain these structures have developed considerable detailed knowledge and skills of a specialist nature. For these reasons it is unlikely that one format will fulfil their diverse requirements.

There is a variety of types of metadata. There is traditional descriptive information of the kind found in library catalogues, which typically includes such attributes as author, title, some indication of intellectual content and so on. There is information that might help a client application make a decision based on format (where certain local browser equipment is available) or on location (to save bandwidth). There are different types of user: a user as customer wishes to know the terms under which an object is available; a user as researcher may wish to have some extended documentation about a particular resource, its provenance for example. There are different types of resource. Some resources may have a fugitive existence, existing to satisfy some temporary need and only ever minimally described if at all; some are important and valuable scholarly or commercial resources, where the value of extensive description is recognised. Some resources may be simple; some may be complex in various ways. There will be many different information providers, some commercial ‘yellow pages’ type services, some scholarly or research-oriented services, in different organisational configurations with different target audiences and products. Metadata may be closely coupled with the object it describes as an intrinsic part of its composition; or it may have no intrinsic link with it at all. And so on ...

Thus, the nature of the problem to be solved suggests a variety of solutions. In the following sections we examine some characteristics of the environment in which network information of interest to European researchers is being created and some of the factors which are influencing the development of metadata services.

Control and the publishing environment

The discipline or control exercised over the production of collections of resources will improve as the web becomes a more mature publishing environment. There will be managed repositories of information objects. Such repositories may be managed by information producing organisations themselves, universities for example, by traditional and 'new' commercial publishers, or by other organisations (the Arts and Humanities Data Service in the UK, for example, or industrial and other research organisations, archives, image libraries, and so on). This is not to suggest that the existing permissive electronic publishing environment will not continue to exist in parallel. One concern of a managed repository will be that its contents are consistently disclosed and that descriptions are promulgated in such a way that potential users, whoever they might be, are alerted to potentially relevant resources in that repository.

Different repositories will have different requirements and priorities. Examples are a social science data archive, a university web site, a commercial publisher's collection of electronic journals, an archival finding list, and so on. Objects on a university web-site may be briefly and simply described. A data archive may need extensive documentation.

A variety of metadata creators and sources

There will be a variety of metadata creators. These fall into three broad categories: 'authors', repository managers, and third party creators. As its importance becomes more apparent, 'authors' are likely to create descriptive metadata: a major incentive for this will be agreement about the use of the META tags in HTML documents for embedding metadata which will be harvested by programs. Descriptive data will be similarly embedded in other objects by those responsible for their creation. Metadata will also be created by repository managers, who have some responsibility for a resource and the data that describes it. Third party creators (including, for example, the information gatways being developed in DESIRE) create metadata for resources which they themselves may not manage or store.

Metadata may sit separately from the resources it describes; in some cases, it may be included as part of the resource. Embedded HTML tags is probably the simplest example of the latter case, but it is common in some of the domain-specific SGML frameworks described in the review section. For example, a TEI header needs to accompany conformat TEI documents. However, independent TEI headers may also exist, which describe documents which may be physically remote.

Metadata, once created may be shared with others. Take for example, author-created metadata embedded in HTML documents. This may be collected by robot or other means. Value will be added to this data at various stages along whatever use chain it traverses: by a local repository manager, by subject-based services like the ones under consideration here, by crawler-based indexing services, by various other intermediary services. These intermediary services might include librarians and others who now invest in current awareness and SDI (selective dissemination of information) services, as well, maybe, as current abstracting and indexing services. Many authors may only provide basic information: typically they will not be conversant with controlled subject descriptor schemes, record all intellectual or formal relationships with other resources, and so on.

A different use chain might be traversed by fuller metadata associated with the scholarly edition of an electronic text, for example. Full documentary metadata would be available to assist in the analysis and use of the text, but a subset might be output to a general purpose discovery service. There might be a link back to the fuller metadata from the shorter record.

A number of factors, including the perceived value of a resource, will determine the relative balance between author-produced, added value and third-party original descriptions in different scenarios. The metadata ecology and economy is still in development.

Structure and fullness

The level of created metadata structure (however it is designed) and the level of intellectual input deemed necessary will depend on the perceived value of the resources and the environment of use.

Webcrawlers tend to describe individual web pages. Newer approaches based on manual description have initially tended to focus on servers, and not describe particular information objects on those servers or the relationships between objects. Most subject information gateways, such as those in the UK eLib project, fall into this category. Neither approach is complete as users are interested in resources at various levels of granularity and aggregation which may not be satisfied by either of these simplified approaches. There also exist a number of emerging approaches specialised for a particular community of users. Quite often, these are rich in terms of content and structure: they are created to represent the objects in a collection and the relationships between them. Examples from the archives, museums, and other communities are given below.

A tripartite division along these lines is further elaborated below. Web indexes based on robot extraction of (currently unstructured) metadata are cheap to create, are automatic. Documentation of a particular collection by specialists is expensive. ‘Information gateway’ services add value through intellectual effort, and are correspondingly expensive. These factors will drive the creation of author-produced metadata and more sophisticated automatic extraction techniques. However, the creation of full, structured metadata will remain expensive, wherever along the use chain that cost falls.

A distributed environment

Programs will collect and manipulate data in a variety of ways, passing it off to other applications in the process. Data may be collected and served up in various environments, converted on the fly to appropriate formats: it may be collected by robot, added to a local 'catalogue', or pulled into a subject-based service. The metadata we have been talking about refers to network information resources. This will need to be integrated at some level with the large, albeit highly fragmented, metadata resource for the print literature. There may also be metadata about people, about courses, about research departments and about other objects. Programs might periodically look for resources that match a particular user profile, might search for people with a particular research interest, and so on.

These developments will take place in a rapidly changing distributed environment in which directory protocols (e.g. whois++, LDAP), search and retrieve protocols (e.g.Z39.50), Harvest, and a variety of other approaches will be deployed. These will be hidden from the user by the web, itself likely to be transformed by the integration of Java and distributed object technologies.

SOME CHARACTERISTICS OF INVESTIGATED METADATA FORMATS

Here we briefly examine characteristics of the metadata formats considered in this study, taking as a framework the broad categories which structure the format descriptions in Part II.

One can suggest an approximate grouping along a metadata spectrum which becomes successively richer in terms of fullness and structure. For purposes of analysis, we propose three bands within this spectrum, which allows us to sketch some shared characterstics across groups of formats. Any one format may not have all the characteristics of the band in which it is placed, but this grouping has proved beneficial in identifying the differences and similarities between formats.

Figure 1: Typology of metadata formats

	Band One	**Band Two**	**Band Three**
Record characteristics	Simple formats	Structured formats	Rich formats
	Proprietary	Emerging standards	International standards
	Full text indexing	Field structure	Elaborate tagging
Record formats	Lycos	Dublin Core	ICPSR
	Altavista	IAFA templates	CIMI
	Yahoo etc	RFC 1807	EAD
		SOIF	TEI
		LDIF	MARC

Environment of use

Band One includes relatively unstructured data, typically automatically extracted from resources and indexed for searching. The data has little explicit semantics and does not support searching by field.

Currently, this data is created by the web crawlers. Many services exist based on such data, and several global services are in heavy use. If a user is looking for a known item, they can be reasonably effective. Because they are global in scope and operate on limited descriptions they are less effective for discovery. A user may find many resources, but may have to sift through them and will miss many potentially relevant resources because they are not indexed with appropriate terms. Nor, in many cases, is the metadata full enough to allow the user to make relevance judgements in advance of actually retrieving the resource. Typically, crawlers are not selective about the resources they index: they often aim for comprehensiveness at some level within their target area, whether that is the world or some part of it. For these reasons, they have some limitations as discovery services. These issues are well known and such services are seeking to enhance the metadata on which they operate: different services have different conventions to allow authors of web pages to include various categories of metadata which can then be collected. There is also some discussion about a common representation for the exchange of such metadata between global indexes and other services, and the harvesting of fuller metadata. We

do not look in detail into such indexes here as they are the subject of a future working paper in the Indexing and Cataloguing component of DESIRE.

Band two includes data which contains full enough description to allow a user to assess the potential utility or interest of a resource without having to retrieve it or connect to it. The data is structured and supports fielded searching. Typically these records are simple enough to be created by non-specialist users, or not to require significant discipline-specific knowledge. Descriptions tend to be of discrete objects and do not capture multiple relationships between objects. Typically, but not essentially, descriptions are manually created, or are manual enhancements of automatically extracted descriptions, and they include a variety of descriptive and other attributes. They may be created to be loaded directly into a discovery service or to be harvested.

Services in this area include OCLC's NetFirst (based on its own internal format) and the UK Electronic Libraries Programme subject-based information gateways (some of which use their own internal format; some use IAFA templates). Often, these services involve some selectivity in what they describe and may have more or less explicit criteria for selection. For these reasons, they may be expensive to create, again driving an interest in author- or publisher- generated description and automatic extraction techniques such as those piloted by Essence as part of the Harvest software.

Our third Band includes fuller descriptive formats which may be used for location and discovery, but also have a role in documenting objects or, very often, collections of objects. Typically, they are associated with research or scholarly activity, require specialist knowledge to create and maintain, and cater for specialist domain-specific requirements. The are expressive enough to capture a variety of relationships at different levels. Developments described below include the Inter-university Consortium for Political and Social Research SGML codebook initiative to describe social science data sets, the Encoding Archive Description, Content Standards for Digital Geospatial Metadata and Computer Interchange of Museum Information.

It should be clear that these are not watertight categories, especially as implementations may vary. GILS and CIMI object descriptions might be considered to be in the middle band for example.

Against this background one can note some trends, especially across the boundaries of these bands. Author or site produced metadata will become more important for many purposes. This may be harvested unselectively, or only from selected sites. An important motivation for this is to overcome some of the deficiencies of current crawlers without a provider incurring the cost of record creation. In some respects, the crawlers will assume some of the characteristics of the middle band.

At the same time, communities using the richer 'documentation' formats will wish to disclose information about their resources to a wider audience. How best to achieve this will have to be worked out: perhaps 'discovery' records will be exported into other systems. These trends suggest that the middle band will become more important as a general-purpose access route, maybe with links to richer domain-specific records in some cases.

Format issues

Metadata formats

There is currently no widely-used standard for data in band one, although amongst implementors of systems based on harvesting of simple metadata there are moves to develop an exchange format based on basic level SOIF. There is also a trend noted above to enhance the data collected by these services in various ways, making them better suited to discovery.

The middle band metadata used in discovery services tends to be based on simple record structures influenced by RFC-822 style attribute-value pairs. Formats here do not contain elaborate internal structure, do not easily represent hierarchical or other aggregated objects, nor, typically, do they express the variety of relationships which might exist between objects. This is usually by design: there is a necessary trade-off between simplicity and expressiveness. Also, their purpose is to be hospitable to the non-specialist description of information objects of different types and from different domains and so is not concerned with the very specific requirements of any one domain. Of the discovery service formats which we examine here, IAFA templates are perhaps the most detailed. There are templates for different types of object (document, user, logical archive, etc.), and there has been some consideration given to 'clusters' of data which are likely to be repeated across records and to variants within records.

There has been some interesting recent discussion about the future direction of the Dublin Core in this context. The Dublin Core is a simple resource description format. It could be extended in two ways. Firstly, it could be extended to accommodate elements which contain other types of metadata: terms and conditions, archival responsibility, administrative metadata and so on. Secondly, it could be designed for resource description of different levels of fullness and within different communities. The IAFA document template is an example of one such format, USMARC another. We would argue that it is undesirable either that there be one single format for resource description or that a single format be indefinitely expanded to accommodate all future requirements. The need to retain a Dublin Core optimised for its target use together with the need to exchange a variety of types of metadata led to the proposed Warwick Framework (which is described in Part II). This is a container architecture for the aggregation of metadata objects and their interchange. However, such an architecture is not yet in place and implementation details are far from clear. It is therefore inevitable that there be a continuing tension between simplicity and the need to provide more expressiveness or functionality.

Although the bulk of the formats in this range follow an attribute-value pair structure, it has been agreed that an SGML DTD will be developed for the Dublin Core. At the 'documentation end' of discovery it is likely that other formats will be found. MARC is a notable one which will be further considered below, but the encoding of choice is now likely to be SGML as in CIMI object descriptions.

Because of some similarity of construction and content across formats in this band, conversion between them, though inevitably lossy, is feasible.

The documentation band contains some very full frameworks for the description of multiple aspects of objects and collections of objects. In some cases, the frameworks describe metadata objects as one type only of information object: they are concerned with 'information content' also. Typically, work is proceeding within an SGML context and the example of the Text Encoding Initiative has been quite influential. Within the social science, museums, archives and geospatial data communities work is progressing on establishing DTDs. These may relate to collection level description, item level description, and allow various levels of aggregation and linkage appropriate to the domain. They cater for a very full range of attributes appropriate to documenting data sets or other resources. These can be distinguished from the range in the middle band by fullness (they go into more detail), structure (they contain richer structuring devices), and specialism (they may be specific to the relevant domain).

It seem likely that specialist users will want to search such data directly, but that to make data more visible to more general 'discovery' tools, there may be export of data in some of what we have called 'discovery' formats. Indeed, the Dublin Core has been explicitly positioned as a basis for semantic interoperability across richer formats, although it has not been widely used in this context.

Protocol issues

Middle band discovery services are being delivered through emerging distributed searching and directory approaches on the Internet, notably whois++, LDAP, and Dienst. There is some use of Z39.50 also, notably for GILS.

Band three documentation approaches are in early stages. However, there has been some discussion of using Z39.50 for search and retrieve in several cases. In particular, there has been some interest in the Z39.50 profile for access to digital collections <URL:http://lcweb.loc.gov/Z3950/agency/profiles/collections.html>.

Implementations

Standards-based resource discovery services are also in early stages. Examination of the descriptions collected in Part II of this report will show that many formats are still under development or are not widely implemented.

In Band 3, the 'documentation category', in particular, communities of users are working towards consensus and in some cases robust interoperating implementations are some time away.

In Band 2, the 'discovery category', IAFA/whois++ templates are in use in several projects, and are deployed in whois++ directory services. Dublin Core is being piloted in several projects, but an agreed syntax is only now being defined. RFC-1807 is used within the NCSTRL project <URL:http://www.ncstrl.org>. SOIF is widely used as the internal format for Harvest, but there is no agreed 'content' definitions. LDIF is in a similar position, lacking an agreed set of schema for resource description. LDIF and SOIF have attracted much interest as a result

of Netscape's decision to base its directory server and catalog server products on LDAP and Harvest respectively.

Of course, an exception to this shallowness of implementation experience is MARC and MARC-like formats. There are many millions of MARC records worldwide, and there are elaborate organisational and technical infrastructures in place for creating and sharing them. MARC is special in this context because of its long established use and its centrality in the library community for describing print resources. There are several initiatives attempting to integrate descriptions of print and electronic resources through the use of MARC and some of these are described in the entries for Pica+ (not a MARC format, but a close analogue), MARC, UKMARC and USMARC. Some library organisations have a vested interest in using MARC for the description of network resources as it simplifies meshing existing sytems with new requirements. It should be noted that MARC records are only standardised at a certain level. ISO 2709 standardises a physical encoding for records. However, each national or other format defines its own set of designators and different rules determine the format of the data content. Several national formats have made changes to accommodate electronic resources. It is likely that conversion into and out of MARC will always be an issue that may have to be addressed by service providers in some contexts.

The majority of existing Z39.50 applications involve searching of MARC based resources. However, this may gradually change as other profiles are introduced.

CONCLUSION

It is clear then, that the organisational, service and technology contexts of resource discovery services are not stable and that risk-free selection of approaches or confident prediction of future scenarios is not possible. Importantly, there is no single driving agency for these developments. Vested interests, competitive advantage, integration with legacy systems or custom and practice will always mean that there are differences of approach.

Choices made within DESIRE must acknowledge this wider context of change.

A number of crosswalks (high level mapping tables for conversion) are now available and can be referenced at UKOLN's metadata web pages at <URL: http://www.ukoln.ac.uk/metadata/interoperability/>

PART II - A REVIEW OF METADATA FORMATS

BIBTEX

Environment of use

Documentation

Bibtex is a program originally designed by Oren Patashnik to create bibliographies in conjunction with the LaTeX Document Preparation System. LaTeX, available for most computer systems, is a system for typesetting documents, independent of the output device. (It is based on the TeX typesetting system by Donald Knuth). BibTeX is a separate program that produces the source list for a document, obtaining the information from a bibliographic database. BibTeX is described in *LaTeX: A Document Preparation System (user's guide and reference manual)*, by Leslie Lamport (Addison-Wesley, 2nd ed. 1994). Further documentation: *BibteXing* and *Designing BibTeX Styles*, both by Oren Patashnik, February 8, 1988.
<URL:ftp://ftp.shsu.edu/tex-archive/biblio/bibtex/distribs/doc/>

Constituency of use

LaTeX is used in scientific and academic communities, and in industry. Scientists use it to send their papers electronically to colleagues over the world. For this reason it is used inhouse and by many STM publishers.

Ease of creation

Experience with the LaTeX system is required to use BibTeX.

Progress towards international standardisation

The first widely available version of LaTeX (2.09) appeared in 1985. Since then various non-standard enhancements were made, which would not work properly at all sites. A new version (2e) was released in 1994.

Format issues

Designation

LaTeX allows for a variety of bibliography styles. The LaTeX input file must contain a \bibliography command whose argument specifies one or more files that contain the database (bib files), and a \bibliographystyle command, that specifies the format of the source list.

The standard bibliography styles are the following (but a lot of other styles are available):

- plain: formatted more or less as suggested by van Leunen in *A Handbook for Scholars* (Oxford Univ. Press, revised ed. 1992). Entries are sorted alphabetically and are labeled with numbers.
- unsrt: the same as plain except that entries appear in order of their first citation.
- alpha: the same as plain except that source labels, formed from the author's name and the year of publication are used.
- abbrv: the same as plain except that entries are more compact because first names, months names, and journal names are abbreviated.

BibTeX provides entry types for almost any kind of reference within a bibliography. Each entry has its own set of fields, divided into three classes: required, optional and ignored (the last for information that shouldn't get into the bibliography).

In the standard bibliography styles the following entries may be used:

- article
- book

- booklet
- conference
- inbook (part of a book)
- incollection (part of book with is own title)
- inproceedings (article in conference proceedings)
- manual (technical documentation)
- masterthesis
- misc
- phdthesis (Ph.D. thesis)
- proceedings
- techreport
- unpublished

Content

Basic descriptive elements

The following is a list of all fields recognized by the standard bibliography styles.

- address: usually address of publisher or other type of institution
- annote: annotation
- author
- booktitle
- chapter
- crossref: the database key of the entry being cross referenced
- edition
- editor
- howpublished
- institution: the sponsoring institution of a technical report
- journal: (abbreviations are provided for many journals)
- key: used for alphabetizing, cross referencing, and creating a label when the 'author' information is missing
- month: the month in which a work was published, or an unpublished one was written
- note: any additional information
- number
- organization
- pages
- publisher
- school
- series
- title
- type
- volume
- year

Subject description

No special fields defined in the standard styles.

URIs

No special fields defined in the standard styles.

Resource format and technical characteristics

No special fields defined in the standard styles.

Host administrative details

No special fields defined in the standard styles.

Administrative metadata

No special fields defined in the standard styles.

Provenance/Source

No special fields defined in the standard styles.

Terms of availability/copyright

No special fields defined in the standard styles.

Rules for the construction of these elements

BibTeX is not designed for use with any specific set of cataloguing rules. For some fields a few (simple) rules are given, such as for the form of the name of the author. The bibliographic style decides how the field content will appear in the bibliography.

Multi-lingual issues

There are commands in LaTeX to generate accents and special symbols used in most western languages. This makes it possible to put bits of non-English text in an English document. They are not adequate for writing a complete document in another language. There is a Babel package which allows the creation of documents in languages other than English as well as multi-language documents.

Ability to represent relationships between objects

By crossreferencing links can be made between different entries.

Fullness

As the primary aim of BibTeX is not to create a bibliographic database that deals with a broad range of bibliographic data, but to format bibliographic references in (scientific) papers, the format is not very extensive. For its purpose the range of fields seems sufficient.

Protocol issues

Not associated with particular protocols.

Implementations

LaTeX is widely used in the scientific communities. Some publishers issue their own macros to enable researches to format their papers according to the standard of the journal.

CATEGORIES FOR THE DESCRIPTION OF WORKS OF ART (CDWA)

Environment of use

Documentation

The *Categories for the Description of Works of Art* were developed by the Art Information Task Force (AITF), sponsored by the Getty Art History Information Program (AHIP) and the College Art Association (CAA).

The *Categories* were released in February 1996, free of charge in both hard copy and in a full hypertext publication in PC and Macintosh versions. Simultaneously a double issue of the journal *Visual Resources* was published devoted to the *Categories* and their use. Plans exist to mount the hypertext document on the Getty website in the near future.

Constituency of use

The *Categories* are developed for the communities that provide and use art information (e.g. museums and archives) and provide a structure for information used to describe works of art and images of them. They focus upon 'movable' objects and their images, including paintings, works on paper, sculpture, ceramics, metalwork, furniture, design, performance art, and so on, from all periods and all geographic regions. The *Categories* can serve three functions: as a mapping document to correlate diverse databases; as a planning document for designing new databases or for extending existing databases; and as a measure against which to evaluate automated tools.

The *Categories* intitiative maintains active liaisons with the CAA's Committee on Electronic Information, the Art Libraries Society of North America (ARLIS/NA), the Visual Resources Association's Data Standards Committee, the Museum Computer Network (MCN), and the Computer Interchange of Museum Information (CIMI) consortium.

Ease of creation

The *Categories* are very extensive and developed for use by art specialists.

Progress towards international standardisation

As the *Categories* have only just been released, the future status is still uncertain. A few test projects have been initiated (e.g. five full cataloging examples from scholars in different specializations have been commisioned and will be put up on the website in the form of hypertext documents) and feedback form constituents is being collected. The fact that the *Categories* have already been mapped to other existing data standards (CIDOC data model, CIDOC MICMO, CHIN data dictionary, ICOM AFRICOM, MDA Spectrum, FDA guide), might influence the adoption of the *Categories* as a standard in the future.

Format issues

Designation

There are 26 main categories, and each category has its own set of subcategories.

Each category and subcategory has been defined using a consistent template. Each category opens with an overall *Definition*, followed by a list of the subcategories. There follows a *Discussion* of the art-historical importance of the information, including its purpose to the researcher, its nature or characteristics, and possible sources for the information. The rubric *Relationships* identifies other categories that contain related information and distinguishes between seemingly similar categories. Under the rubric *Uses* is a discussion of how the information might be applied in research. *Access* describes ways in which researchers might wish to retrieve the information. The subcategory template is the same as the category template, with the addition of *Examples* following the *Definition*, and the rubric *Terminology/Format*, under which applicable controlled vocabularies and other resources are identified.

Content

Basic descriptive elements

The *Categories* are a statement of the intellectual content for a description of a work of art, but do no represent database fields or database structures for managing art information. When building an implementation based on the Categories, an institution needs to determine how to structure the data and what level of specificity best suits its needs.

The main categories are:

- Object/work
- Classification
- Orientation/arrangement
- Titles or names
- State
- Edition
- Measurements
- Materials and techniques
- Facture
- Physical description
- Inscription/Marks
- Condition/Examination history
- Conservation/Treatment history
- Creation
- Ownership/Collecting history
- Copyright/Restrictions
- Styles/Periods/Groups/Movements
- Subject matter
- Context
- Exhibition/Loan history
- Related visual documentation
- Related textual references
- Critical responses
- Cataloging history
- Current location

Subject description

- Classification

  Consists of the following subcategories:

- Term: the specific term or code from a formal classification scheme that has been assigned to a work.
- Remarks: Additional notes or comments pertinent to the classification of a work of art.
- Citations: An identification of the scheme or structure from which the classification term is drawn.

  Other categories relating to content are:

- Subject matter
- Context
- Styles/periods/groups/movements

URIs

No special category specified.

Resource format and technical characteristics

Not specified.

Host administrative details

Not specified.

Administrative metadata

- Cataloging history

Provenance/source

- Ownership/Collecting history

Terms of availability/copyright

- Copyright/Restriction

Defined as: An identification of the individual or group that holds the rights to use, exhibit, or reproduce a work of art, along with an indication of any existing restrictions on its reproduction, exhibition, or use.

Rules for the construction of these elements

For a number of subcategories the use of controlled vocabularies and authority files to provide consistent access to names of people and places and to descriptive terminologies is recommended

Encoding

Work on an SGML DTD is in progress.

Ability to represent relationships between objects

Some categories are provided to contain information about:

- Related works
- Related visual documentation
- Related textual references

Fullness

Full.

Protocol issues

Dependent on the kind of database that is created with the *Categories*.

Implementations

The Inventario del Patrimonio Cultural in Chile has mapped their new database to the *Categories*, and the Hispanic Society in New York is doing the same. The member institutions of MESL (Museum Educational Site Licensing Project) are also taking the Categories into consideration.

The categories are being tested in the context of the CHIO project (see CIMI entry).

Environment of use

Documentation

The Museum Computer Network (MCN) first proposed an initiative to investigate standards for interchange of museum information in 1988. The outcome of this effort was the CIMI Committee, which was active from 1990-92 and produced *A Standards Framework for the Computer Interchange of Museum Information*, by David Bearman and John Perkins (MCN, May 1993. Available from <URL:http://www.cni.org/pub/CIMI/www/framework.html>, and also published as a double issue of the MCN quarterly SPECTRA, vol. 20, no 2 and 3). This framework encompasses interchange protocols, interchange formats, and lower level network and telecommunications building blocks as well as content data standards.

The CIMI Consortium continues the work of the CIMI Committee that resulted in the Standards Framework <URL:http://www.cimi.org/>. The Consortium's current major project is CHIO (Cultural Heritage Information Online). The main aim of the project is to offer in a demonstrator at least 10,000 records of objects and information about Folk Art as a searchable online resource, including the full text of exhibition catalogues, wall texts, as well as images and more traditional museum database records. The CHIO project consists of two parts: *CHIO Structure*, which explores the use of SGML, and *CHIO Access* to explore the utility of Z39.50.

Constituency of use

Museum communities. The CHIO project addresses information needs relating to cultural heritage information in a broad sense: information held by museums, archives and special collections in libraries. As Project CHIO aims to provide information to the average paying museum visitor, an over-academic approach has to be avoided.

Ease of creation

Complex

Progress towards international standardisation

The Standard Framework and especially the choice for SGML and Z39.50 seem to offer ample opportunities to reach a high degree of standardisation within the museum community, but much will depend on the results of CHIO and future projects. The advantage of SGML is that it offers a framework for the encoding of any type of document. Any tags can be used as long as a description of the tags that are used is given at the start of the document, as a DTD (Document Type Definition).

For future developments CIMI has decided that the CHIO DTD should be compatible with HyTime (ISO/IEC 10744), a standard for Interactive Open Hypermedia which is based on SGML and fully SGML-compatible.

At the conclusion of Project CHIO, CIMI expects to see a community endorsed system of encoding museum information using SGML; a methodology for searching texts and collections data using Z39.50; and a demonstration system that will show the power of a standards based approach to electronic interchange.

Format issues

Designation

The museum/cultural heritage communities offer a broad range of information of different types and structures (e.g. structured texts, full-text documents, images). One of the aims of CHIO is to provide different DTDs for different types of information with different requirements, for the marking up of this information with SGML.

The first type of information to be analysed were exhibition catalogues (as a testcase for any text-based museum information resources). The following is based on the DTD for this type of information.

Content

Basic descriptive elements

- Record Summary
- Dates
- Object Title Name
- Document Title
- Editor
- Person Name
- Organization Name
- Place
- Record type

Subject description

- Classification
- Concept
- Event
- Material
- Mark
- Object
- Object Identifier
- Occupation
- Role
- Style/Movement
- Subject
- Topic

URIs

- Not defined.

Resource format and technical characteristics

- Not defined.

Host administrative details

- Organization Name
- Place

Administrative metadata

- Record Type
- Document Source
- CHIO Contributor
- SGML Source File Name
- CHIO Document No.

Provenance/source

- Document Source

Terms of availability/copyright

- Copyright

Rules for the construction of these elements

The CIMI Standard Framework can be implemented at two levels. The first is specification of hardware and software so that it supports the standards defined in the CIMI Standards Framework. This will ensure that the data can be interchanged even if all the institutional meanings cannot be. The second level adresses the standardization of data content (the fields of information), and data values (what goes in the fields).

Designation and encoding

CIMI is using SGML (and will eventually use other standards too like HyTime and Z39.50), to make museum information available.

SGML is used to express the structure and content of exhibition catalogues and as the foundation for a data interchange format for collections records. CIMI has developed DTDs along with stylesheets and a navigator which can be seen in action in the CHIO Demonstrator.

CIMI decided to develop a comprehensive set of museum DTDs, one for each genre of museum information, rather than one over-generalized DTD for all museum information. It was decided that the DTD should allow for all the significant features of the source document to be marked up. For exhibition catalogues and wall texts the TEI Lite DTD (a cut-down version of the TEI prose DTD) offered a good starting point. Based on the CIDOC Data Model the TEI Lite framework was modified and extended with access point tags to support the access needs of the CHIO project (this was done following the TEI Guidelines).

It was agreed that database records have different requirements and need a separate DTD.

Ability to represent relationships between objects

The need for a robust linking mechanism to connect the parts of the distributed public information resource is acknowledged. It will depend on future developments relating to FPI (formal public identifiers). For more information on the proposals for the FPI resolution scheme see *James K. Tauber, FPI-URN Resolution on the Internet: Notes* <URL:http://entmp.org/fpi-urn>.

Protocol issues

The foundation of the original MCN Standards Framework was the Open Systems Environment (OSE) Reference Model. CIMI addresses the interchange aspects of the OSE model that includes how data is represented, how various data types are identified, and how data content objects are presented.

Transport services can be provided by OSI or an appropriate alternative such as TCP/IP. CIMI uses high level OSI protocols such as FTAM for file transfer, X.400 and X.500 for messaging and directory services, and ISO 9040/41 for terminal access but these can be run over any appropriate lower layer transport protocols. CIMI recommends EDI for business transactions and ISO 10162/10163 for information retrieval. At first no consensus was reached on one standard for the building of collections databases and reference files, only rationales were given for using either ISO 2709 MARC, ISO 8879 SGML, or ISO 8824 ASN.1. But since then a positive choice has been made for ISO 8879 SGML.

The CHIO Structure project database can be accessed over the Internet using generic WWW browsers. CIMI is developing the CIMI Profile: Z39.50 Application Profile Specification which is a specification of what features of Z39.50 should be implemented in what way to provide the required functionality.This profile has now been issued as a draft for comment <URL:http://www.cni.org//pub/gils/profile/final.report/>.

The Profile is reaching a point of stability and has the consensus of the working group that has been working on it since September 1995. After the comment period, and necessary revisions, the revised version will become the

initial version of the draft Profile upon which implementors will build support for search and retrieval in Project CHIO.

The CIMI Profile requires the following basic services of Z39.50-1995:

- Initialization, including ID Authentication
- Search, for searching the CHIO Information Resource
- Present, for retrieval of information objects
- Access Control, for handling copyright information

The CIMI Profile will include a CIMI Attribute Set, which will enable the expressions of queries for searching cultural heritage museum information resources. The Profile will use the Z39.50 Generic Record Syntax (GRS) for packaging retrieved records for presentation to the client.

The initial Z39.50 implementations of the CIMI Profile will support access to a demonstration CHIO Information Resource. The CHIO Information Resource can be modeled as a digital library comprised of hierarchical, distributed collections of digital information. The CHIO Information Resource consists of a number of physical and/or logical datastores of museum information, and the datastores may consist of one or more databases. A user may search the CHIO Information Resource to retrieve digital information objects in several possible data types (SGML, MARC, other structured records, Image, Audio, Video).

The CIMI Profile is being developed as a companion profile to the Z39.50 Profile for Access to Digital Collections (the Collections Profile) <URL:http://lcweb.loc.gov/Z3950/agency/profiles/collections.html>. The focus of the Collections Profile is the access and navigation of digital collections. The focus of the CIMI Profile is the search and retrieval of specific information resources contained in the digital collections.

Implementations

The CIMI Z39.50 Profile which is being developed is intended to support a demonstration project rather than working implementations. A pragmatic approach has been agreed so that implementation experiences will guide subsequent extensions, and existing implementations that will serve up museum information for Project CHIO will be the basis on which a number of Profile decisions will rest (e.g. attribute sets, element sets, record syntax, etc.). During the period Summer 1996 through Spring/Summer 1997, interoperability testing will provide input to revisions and changes to the CIMI Profile. By the end of the Project CHIO (Fall 1997), a final, revised CIMI Profile will be completed.

Note that the CHIO demonstartor requires the use of an SGML browser such as Panorama, in addtion to a generic web browser, in order to view the SGML encoded documents.

DUBLIN CORE

Note: see also the entry for Warwick Framework

Environment of use

Documentation

'Dublin Core' is shorthand for the Dublin Metadata Core Element Set which is a core list of metadata elements agreed at the OCLC/NCSA Metadata Workshop in March 1995. The workshop report forms the documentation for the Dublin Core element set. (Stuart Weibel, Jean Miller, Ron Daniel. *OCLC/NCSA metadata workshop report*. OCLC, March 1995. <URL:http://www.oclc.org:5046/conferences/metadata/dublin_core_report.html>)

Constituency of use

The workshop was organised by OCLC and the National Centre for Supercomputer Applications (NCSA) to progress development of a metadata record to describe networked electronic information. This workshop followed on from joint meetings and discussions of the American Library Association. The workshop brought together a range of interested parties from different professional backgrounds and subject disciplines, all of

whom had been involved with metadata issues. The motivation progressing Dublin Core has been to reach a consensus among stakeholders on a minimal resource description which can be used for the benefit of all involved in the creation, search and retrieval of electronic resources. There has been high commitment and involvement from a range of professions (publishers, computer specialists, librarians and information workers) and sectors (library utilities, software producers, service providers, libraries).

The Dublin Core is positioned as a simple information resource description. However, importantly it also aims to provide a basis for semantic interoperability between other, probably more complicated, formats. A third target use is to provide the basis for resource-embedded description, initially with HTML documents.

Ease of creation

The objective of Dublin Core is to define a simple set of data elements so that authors and publishers of Internet documents could create their own metadata records with no extensive training. The Dublin Core approach is to have the level of bibliographic control midway between the detailed approaches of MARC and 'structured' TEI, and the automatic indexing of locator services such as Lycos. It is acknowledged that the Dublin Core is a minimal set, and that many 'publishers' or metadata producers may wish to augment this simple set with more specialised data.

Progress towards international standardisation

Initial attempts to include consideration of Dublin core elements as part of an IETF working group were not taken forward, on the grounds that the content of metadata records is outside the scope of IETF standards. However the Dublin core elements have been considered by USMARC as central to their development of the USMARC record so the impact has already been seen in the formation of other metadata.

Ambitions to actualise Dublin Core were carried forward by a second international workshop which took place in the UK at the University of Warwick in April 1996 sponsored by UKOLN and OCLC. This workshop looked at the implementation of Dublin Core and the requirements for extensibility, change control and dissemination. The need for a registration agency was discussed at this meeting.

Format issues

Designation and encoding

The Dublin Core is a set of elements that can be used to describe a resource but there was initially no attempt to prescribe an encoding method or record structure. During the first Dublin Core workshop there was an explicit decision taken not to define syntax at this stage.

However certain principles were established for further development of the element set. Of particular relevance to encoding and designation are the principles of

- extensibility: the core set can be extended with further elements to describe intrinsic data of particular relevance to a particular community
- optionality: all elements are optional
- repeatability: all elements are repeatable
- modifiability***:*** any element can be modified by one or more qualifiers

The sanctioning of qualifiers is of particular note as it is an attempt to bridge the gap between casual and sophisticated use. Qualifiers can be of two very different types: some indicating external schemes to be applied to processing e.g. OtherAgent(scheme=TEI), some specifying more precise information about the attribute, in effect sub-dividing the element name e.g. OtherAgent(role=editor). If a scheme qualifier is used then this means the syntax of that scheme must be applied to the data in that element. So Author (scheme=USMARC) fields will contain data embedded with USMARC tags and sub-field markers, and OtherAgent (scheme=TEI) elements will contain data with TEI mark-up tags embedded. Potentially widespread use of qualifiers could cause severe problems with interoperability.

At the Warwick workshop a decision was taken to develop a concrete syntax for the Dublin Core in the form of an SGML DTD. (The proposed syntax is described in: Lou Burnard, Eric Miller, Liam Quin, C.M. Sperberg-

McQueen, *A Syntax for Dublin Core Metadata: Recommendations from the Second Metadata Workshop* <URL:http://users.ox.ac.uk/~lou/wip/metadata.syntax.html>).

Content

Basic descriptive elements

The core element set includes the following bibliographic data elements:

- Title (name of the object)
- Author (person(s) primarily responsible for intellectual content)
- Publisher (agent or agency responsible for making the object available)
- OtherAgent (person(s) such as editors or transcribers, who have made other significant intellectual contributions to the work)
- Date (date of publication)
- ObjectType (genre of the object such as novel, poem, dictionary)
- Language (language of the intellectual content)

The Author element name does not distinguish the form of author (personal/corporate/meeting). Similarly the OtherAgent element name does not express the precise role of the other agent. It would be possible to use qualifiers to make these more precise distinctions, but the Dublin Core documentation does not attempt to make comprehensive recommendations. Suggested qualifiers are:

Author(scheme=USMARC)=100 1 Doyle, Conan $c Sir, $d 1859-1930

OtherAgent(role=editor)=Weibel,Stuart L.

As soon as such qualifiers are used the complexity of processing the data, and the difficulties for interoperability, will increase.

Subject description

The core element set includes the data elements:

- Subject (topic addressed by the work)
- Coverage (the spatial and temporal characteristics of the object)

The subject element can be used for headings controlled by a known classification scheme indicated in the qualifier, or can contain free text. The Coverage element allows spatial or temporal data to be included for geospatial data. This data might be in unstructured form or in a format governed by a known scheme e.g.

Coverage(type=spatial)=Atlantic ocean

Coverage(type=spatial,scheme=LATLONG)=West=180,East=180,North=90,South=90

URIs

The core element set includes the data element:

- Identifier (string or number used to uniquely identify the object)

The data in this element could be an identifier conforming to an internationally recognised scheme (e.g. URL, ISBN) or it could be a local, privately administered number (e.g. university technical report number). The qualifier would need to be used to make the identifier generally useful.

Resource format and technical characteristics

The core element set includes the data element:

- Form (the data representation of the object such as Postscript file or windows executable file)

A constraint on the design of the Dublin Core, accepted by the workshop participants, was that the aim of the element set is to describe ‘document like objects’ (DLOs).

Administrative metadata

No administrative data is included in the Dublin Core set. A principle of intrinsically was established at the workshop which constrained the set to only include elements describing the intrinsic properties of the object. It would seem essential for any implementation of Dublin Core to include in a record such information as the record identification, record creation date, etc.

Provenance/source

The core element set includes the data element:

- Source (objects, either print or electronic, from which the resource is derived)

This element could be used to link different versions of an object which have the same intellectual content, whereas the relation element would be used to link objects with a different intellectual content.

Host administrative details/Terms of availability/copyright

An agreed constraint on Dublin Core is that extrinsic data such as cost and details of access methods would be excluded from the element set. It was accepted that only elements for resource discovery would be included, not those elements specific to retrieval and request.

Other comments

At the Warwick Workshop it was decided that content-wise the Dublin Core should remain more or less as it was. It should not be indefinitely extended to encompass the variety of current and future metadata requirements.

Ability to represent relationships between objects

The core element set includes the data element:

- Relation (relationship to other objects)

This element describes relationships to other objects with different intellectual content. It allows for a variety of relationships to be identified by use of the qualifier mechanism. Specification of a relationship would require use of at least two qualifiers, e.g.

Relation (type=ContainedIn) (identifier=URL) =http://www.ukoln.bath.ac.uk/metareview.html

Multi-lingual issues

The core element set includes the data element:

- Language (language of the intellectual content)

The problems of use of non-ASCII characters within the record were deliberately not addressed.

Fullness

The fullness of Dublin Core is low, by design. The attempt to compromise with sophisticated use by the qualifier mechanism could potentially lead to highly complex, much fuller records.

Conversion to other formats

MARBI Discussion Paper No 86 (Mapping the Dublin Core elements to USMARC, Library of Congress, May 5, 1995 <URL:gopher://marvel.loc.gov:70/00/.listarch/usmarc/dp86.doc>) looks at options and problems in matching Dublin Core to USMARC. Because Dublin Core elements are less specific than MARC, some fields cannot be sufficiently identified to tag them correctly. For example the author field in MARC is identified as being personal or corporate name, whereas Dublin Core does not make this differentiation.

Other crosswalks are available:

From Dublin Core to USMARC by Rebecca Guenther / Network Development and MARC Standards Office (Library of Congress) <URL:http://lcweb.loc.gov/marc/dccross.html>

From Dublin Core to EAD/GILS/USMARC - by Eric Miller (OCLC).
<URL:http://www.oclc.org:5046/~emiller/DC/crosswalk.html>

From Dublin Core to IAFA/ROADS templates - by Michael Day (UKOLN).
<URL:http://www.ukoln.ac.uk/metadata/interoperability/dc_iafa.html>

From Dublin Core to Z39.50 tag set G - by Ray Denneberg (Library of Congress) - mail to Meta2 list, Feb. 1997.
<URL:http://www.roads.lut.ac.uk/lists/meta2/0733.html>

Rules for construction of these elements

No formulation of rules

Protocol issues

Not yet applicable.

Implementations

There have been a few early implementations of Dublin Core.

National Document and Information Service <URL:http://www.nla.gov.au/2/NDIS/NDISintro.html>: this is a joint project between the National Libraries of Australia and New Zealand. Within this project the Dublin Core elements have been used as the core search attributes for their records, in effect the intersection between their various databases. There has been flexibility in the use of semantics with mapping of other 'search fields' to the Dublin Core set.

DSTC <URL:http://www.dstc.edu.au/>in Australia is using the Dublin Core in the Research Data Network Co-operative Research Centre project for resource discovery.

ENCODING ARCHIVAL DESCRIPTION (EAD)

Environment of use

Documentation

EAD consists of an SGML DTD, a Tag Library, Guidelines for its use, and examples. The Library of Congress Network Development/MARC Standards Office acts as the maintenance agency and the Society of American Archivists is the owner of the emerging standard, and is responsible through a committee representing the archival community for ongoing oversight and development.

(Note: at the time of writing the Beta EAD DTD and Tag Library are available at the Library of Congress site: <URL:ftp://ftp.loc.gov/pub/ead/> or from the current EAD web site at Berkeley: <URL:ftp://library.berkeley.edu/pub/sgml/ead/>).

Constituency of use

The EAD has been developed for use with archives and manuscripts collections. It was motivated by a desire to provide an enduring standard for machine representation of archival description and to facilitate uniform network access to archive and manuscript library collections. While MARC records provide summary description and access, EAD is intended to provide detailed description and access. The two descriptive methods are intended to be complementary, with MARC records providing summary representation of collections in bibliographic databases that lead to the detailed EAD-based finding aids (i.e. detailed catalogues). EAD provides a structure for describing archive and library finding aids and is primarily intended for inventories and registers. It accommodates registers and inventories of any length describing the full range of archival holdings in various media.

(Note: US usage is ‘finding aids’, while UK usage calls similar descriptions ‘detailed catalogues’.)

Ease of creation

The EAD provides an apparatus for full, hierarchical description and is designed for use by those with a knowledge of collections and archival practice.

Progress towards international standardisation

EAD is not an international standard, though is receiving international interest. The Library of Congress Network Development/MARC Standards Office acts as a maintenance agency for EAD. The Society of American Archivists will retain ongoing oversight for development. The Beta version of EAD and an electronic version of the guidelines are scheduled to appear in late Summer or Autumn of 1996. Discussions concerning the internationalisation of EAD and compliance with International Standard Archival Description (ISAD(G)) are underway.

Other comments

EAD recognises that the TEI guidelines and USMARC are related and complementary. The data model includes a finding aid header which is based on, and thus similar to, the TEI header. (*Encoding standard for electronic finding aids: a report by the Bentley Team for Encoded Archival Description Development*. n.d.).

Format issues

Content

Basic descriptive elements

An overview of high-level descriptive elements is given here. Lower-level descriptive elements and some other elements are not listed. This list reflects the standing of the DTD as of August 1996, and may be changed in later versions.

<ead>
  <eadheader>
    <filedesc>
      <titlestmt>
        <titleproper>
        <subtitle>
        <author>
        <sponsor>
      <editionstmt>
      <publicationstmt>
        <date>
        <publisher>
        <address>
      <seriesstmt>
      <notestmt>
    <profiledesc>

<creation>

<langusage>

<revisiondesc>

<runningft>

<frontmatter>

<findaid>

<archdesc> (Archival Description)

<did> (Descriptive Identification)

<note>

<origination>

<physdesc>

<repository>

<unitdate>

<unitid>

<unitloc>

<unittitle>

<admininfo>

<accessrestrict>

<acqinfo>

<altformavail>

<appraisal>

<custodhist>

<note>

<prefercite>

<processinfo>

<userestrict>

<arrangement>

<bioghist>

<chronlist>

<controlaccess>

<corpname>

<famname>

<genreform>

<geogname>

<name>

<occupation>

<persname>

<*subject*>
<*note*>
<*odd*> *(Other Descriptive Data)*
<*organization*>
<*scopecontent*>
<*arrangement*>
<*organization*>
<*dsc*> *(Description of Subordinate Components)*
<*c*> *(Component)*
<*did*>
<*admininfo*>
<*arrangement*>
<*bioghist*>
<*controlaccess*>
<*note*>
<*odd*>
<*organization*>
<*scopecontent*>
<*dsc*>
<*c*>
<*add*> *(Adjunct to Descriptive Data)*
<*bibliography*>
<*fileplan*>
<*index*>
<*relatedmaterial*>
<*separatedmaterial*>
</ead>

Subject description

There is a variety of elements including <subject>, but <scopecontent> is also important.

URI

The finding aid itself has an <eadid> element for uniform representation of a unique identification of the archival collection. In addition, the <unitid> is to be used for unique identification of the archival collection. Emerging IAAD(G) naming/ID conventions need to be reconciled and developed in light of international Internet standards. Finally, the element <dao> or Digital Archival Object for original digital and digital representations of archival material will use SGML formal public indentifiers that will be mapped, using SGML standards, and depending upon the method chosen, to emerging Internet standards: PURL, CNRI handles, and when available other URIs.

(Note: Berkeley and Library of Congress are experimenting with respect to pointing. LC seems to be going with CNRI handles, Berkeley with PURLs.)

Resource format and technical characteristics

The <physdesc> element is used to describe the physical characteristics of the object being described in the <ead> (e.g. a letter written by Mark Twain).

Where a digital representation of the file exists, physical characteristics of the representation (file size, format, information documenting the capture process, etc) will reside in the header of the digital representation file, or if it is maintained separately, in a separate metadata format and syntax (e.g. a digital representation of a letter written by Mark Twain; with separate physical characteristics and capture information on each page-image).

Host administrative details

DTD has elements for repository/authoritative agency. <eadid> has an attribute for explicit representation of authoritative/host institution when this is not in a machine-parsable form in the data in the element itself.

Administrative metadata

Carried in the <eadheader> subelements <filedesc> and <profiledesc>. The element and subelements are based on the TEI header, though simpler.

Provenance/source

<repository>

Terms of availability/copyright

<accessrestrict> and <userestrict>

Designation and Encoding

The EAD has been implemented as an SGML DTD.

Multi-lingual issues

The DTD invokes ISO standard entity sets. Full details from the LC Network Development/MARC standards Office. Usage is compatible with mapping to Unicode when there is sufficient software support for it.

Fullness

Less full than MARC and TEI by design, but still a relatively rich format.

Protocol issues

EAD does not prescribe any search or transport protocols.

Implementations

There are major projects at the Library of Congress and the following US universities: Berkeley, Yale, Harvard, Duke, Stanford, and Virginia. There are also several projects initiated by the National Endowment for the Humanities and also some other projects.

Environment of Use

Constituency

The Engineering Electronic Library (EELS) is a co-operative project of The Swedish Universities of Technology Libraries to provide an information system for quality assessed information resources on the Internet. (see EEVL for a subject based engineering service based in the UK). The metadata format that EELS is using is specific to the project (there does not appear to be any international standard format being used within the Engineering community).

Documentation

The EELS service can be found at <URL:http://www.ub2.lu.se/eel/eelhome.html>

Ease of creation

So far the format has only been used by librarians and information specialists.

Progress toward international standardisation

EELS are currently using a home grown format, however as part of the DESIRE project they are investigating changing to a more standard format (probably either the Dublin Core or IAFA) in the near future.

Format Issues

Content

The EELS template consists of 11 attributes not all of which are displayed to the user.

Basic descriptive elements

- Title (in original language).

Subject description

Assigns classification codes (maximum of 4) and indexing terms (maximum of 10). Uses Engineering Information (Ei) Inc's classification system and thesaurus for the subject description, in order to allow combined searching with Ei's Compendex database, which indexes the printed literature.

URIs

The URL is recorded and used to construct the link to the resource.

Resource format and technical characteristics

Publication types and formats are taken from the following list:

- Archive
- Bibliography
- Book catalogue
- Conference announcement
- Electronic conference
- Electronic journal
- FAQ
- Library catalogue

- Organizational home page
- Publication list
- Reference data
- Table of contents
- Bibliographic database
- Bulletin board (use Electronic conference as well)
- Dictionary
- Directory
- Full-text database
- Image database
- Numeric database
- Properties database
- Software database
- Statistical database
- Time-series database
- Transactional database

Host administrative details
None

Administrative metadata
Editors e-mail address (not used in the display format) and an internal annotation field.

Provenance
Often given in the title or annotation.

Terms of availability
Any login information needed to use the resource is recorded.

Designation
Information is stored using a simple attribute/value pair scheme.

Multilingual issues
The country of origin is recorded and the title of the resource is recorded in the original language. Information is coded in ISO Latin-1.

Ability to represent relationships between objects
None.

Fullness
Medium.

Rules for formulation of data element content
There are no explicit rules.

Protocol Issues

The records are searchable using WAIS (Wide Area Information Server) .

THE EEVL METADATA FORMAT

Environment of Use

Constituency of use

The Edinburgh Engineering Virtual Library (EEVL) is a project funded by eLib (the Electronic Libraries Programme) in the UK to provide an Internet gateway to quality information resources in Engineering, (see EELS for an engineering subject service based in Sweden).Whilst the metadata format that EEVL is using is specific to the project there does not appear to be an alternative international standard format being used within the Engineering community.

Documentation

Information about the EEVL project can be found at <URL:http://www.eevl.ac.uk>. An example of a completed record can be found in <URL:http://www.eevl.ac.uk/pub3.html>.

Ease of creation

The format consists of a template with twenty two attributes that have been chosen specifically to describe networked resources. The creation of records for the project is performed via a WWW interface and many of the administrative fields are assigned automatically.

Progress toward international standardisation

The format was developed in-house for use by the EEVL project. However a conscious effort was made to ensure that most of the fields map directly over to the IAFA format to allow interoperability with other eLib subject gateways using IAFA within the ROADS software.

Format Issues

Content

The format of the template is a simple ASCII record of twenty two attribute/value pairs.

Basic descriptive elements

Bibliographic type elements consist of:

- Title
- Alternative title
- ISSN/ISBN
- Keywords
- Description

There is a subject descriptor field with a selection of descriptors based loosely around the Ei (Engineering Information Inc) classification scheme (multiple types are supported).

URIs

There is a specific field for URL. This is a multi line text box that allows multiple URLs to be input. Where multiple URLs exist however there is no facility for matching a URL to other details such as authentication or contact.

Resource format and technical characteristics

The resource type options are:

- Information server- Higher Education
- Information server- Society/Institution
- Information server- Commercial
- Information server- Governmental
- Library Catalog
- E-journal/Newsletter
- Database/Databank
- Resource Guide/Directory
- Training Materials
- Reference
- Conference/Meeting Announcements
- Recruitment/Employment
- Patents/Standards
- Mailing/Discussion List
- Research Project/Centre
- Software (where freely available)

There is no provision for recording technical characteristics, although there is a free text field for authentication details.

Host administrative details

There is a free text field for a contact email address for the person/organisation responsible for the resource.

Administrative metadata

The administrative information about the metadata is automatically assigned and includes:

- Handle - unique identifier for the template
- Template-creator
- Date created
- Date last modified
- Last modified by

Provenance

None

Terms of availability

A free text field for registration details for the resource if general access is not available.

Designation

The template uses a simple attribute/value pair scheme.

Multilingual issues

No specific provision for other languages.

Ability to represent relationships between objects

None

Fullness

The template covers descriptive elements of a resource well, however it does not deal with the concept of variants of a resource - e.g. the same resource being available from different sites and having different administrative information. It also not able to represent relationships between resources.

Protocol Issues

Planning to move to whois++.

Implementations

The format is designed specifically for the EEVL project. The project is currently in a pilot phase and the service is being tested by six UK universities.

FGDC - CONTENT STANDARDS FOR DIGITAL GEOSPATIAL METADATA

Environment of Use

Constituency of use

The Federal Geographic Data Committee (FGDC) initiated work in June 1992 on a common set of terminology and definitions for the documentation of geospatial data. The resulting standard was approved by the committee in June 1994 as the *Content Standards for Digital Geospatial Metadata*. The name of the format is strictly speaking the Content Standards for Digital Geospatial Metadata (CSDGM) however it is more commonly referred to as the FGDC standard and will be referred to as such throughout this review.

The collection of geospatial metadata was mandated in the US by Executive Order 12906, *Co-ordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure* in April 1994 (Executive Order 12906. <URL:http://www.fdgc.gov/execord.html>). The order instructs federal agencies to document new geospatial data beginning in 1995 and to provide the metadata to the public through a National Geospatial Clearinghouse. Geospatial data prepared before January 1995 only requires identification and contact information. Responsibility for the preparation of metadata lies with the agency or the source of data.

The National Geospatial Data Clearinghouse is comprised of software and institutions to facilitate the discovery, evaluation and downloading of digital geospatial data and to implement the FGDC standard.

Documentation

Documentation details for the standard can be found at the FGDC web site <URL:http://www.fgdc.gov/Metadata/metahome.html>. The current approving authority for the standard is the FGDC although there are plans to submit it to the Department of Commerce for approval as a Federal Information Processing Standard (FIPS).

Ease of creation

The FGDC standard is a complex format with over 300 data elements. A number of tools have been developed by different agencies to assist with the creation of metadata records, a selection of them are available from <URL:http://www.fgdc.gov/clearinghouse/mitre/task2/tools.html>. The FGDC Clearinghouse Working Group sponsor orientation and training sessions on implementation of the metadata standard and clearinghouse procedures.

Progress toward international standardisation

The FGDC standard was put forward at the first meeting of the International Standards Organization Technical Committee 211 (ISO/TC211) in December 1995. The committee are working on creating an international standard for geographic information (due late 1998). Modifications to the FGDC metadata standard are being run in co-ordination with ISO and FGDC Standards Processes and will enter public review in the next few months, to be implemented by July 1997. (Summary of Actions - FDGC Clearinghouse Working Group, April 4 1996. <URL:http://fgdc.er.usgs.gov/clearinghouse/496/summary.html>).

The FGDC standard has also been mapped to many other existing standards e.g. DIF (the NASA Directory Interchange Format), GILS (Government Information Locator Service), USMARC and the Dublin Core.

Other comments

The standard provides specifications for the information content of metadata for digital geospatial metadata. It is designed to allow a prospective user to determine the availability, fitness for use and means to access the data. However it does not specify "the means by which this information is organised in a computer system or in a data transfer, nor the means by which this information is transmitted, communicated or presented to the user" (Overview: FGDC metadata content standard, June 8 1994. <URL:http://geochange.er.usgs.gov/pub/tools/metadata/standard/overview.html>). All implementation details for the standard are part of the responsibilities of the National Clearinghouse and are dealt with in *Draft Implementation Methods for Access to Digital Geospatial Metadata - Normative Annex D* (ISO/TC211/WG3/WI15/N001 <URL:http://www.fgdc.gov/clearinghouse/annex.html>).

Format Issues

Content

The standard consists of 7 main sections of metadata and 3 'support' sections:

- Section 1: Identification information
- Section 2: Data quality information
- Section 3: Spatial data organisation information
- Section 4: Spatial reference organisation information
- Section 5: Entity and attribute information
- Section 6: Distribution information
- Section 7: Metadata reference information

Support Sections

- Section 8: Citation information
- Section 9: Time period of content information
- Section 10: Contact information

Within these sections elements are specified as mandatory, mandatory if applicable and optional. The only two sections to be deemed mandatory as a whole are Section 1 (the identification information) and section seven (the metadata reference information). The support sections do not stand alone but contain information that is referenced more than once in the seven main sections.

Basic descriptive elements

Basic bibliographic elements of a data set are recorded in Section 1 (Identification information). These include:

- title
- citation
- publisher
- description
- abstract

Subject description

The subject coverage of a data set is described in Section 1 (Identification information) under ‘theme keywords’. A draft on-line thesaurus of geospatial keywords from the NASA Master Directory is being provided by the Clearinghouse.

URIs

Annex D specifies an element ‘on-line linkage’ in the format of a Uniform Resource Locator.

Resource format and technical characteristics

The format and technical characteristics of a data set are recorded in Section 6 (Distribution information):

- digital form
- digital transfer information

Host administrative details

These are provided in section 6 (Distribution information):

- contact organisation
- contact position
- contact address
- contact telephone

Administrative metadata

Section 7 is devoted entirely to administrative metadata:

- metadata date
- metadata contact
- metadata standard name - (i.e. FGDC)
- metadata standard version

Provenance/Source

The source of a data set is provided in the support Section 8 (Citation information):

- originator

Terms of availability

The terms of availability are covered in Section 6 (Distribution information) these include:

- distribution liability
- digital form
- digital transfer information
- fees

Designation

Uses attribute/value scheme. With the adoption of SGML the attribute naming will use an 8 character tag scheme <URL:http://www.fgdc.gov/clearinghouse/tag8names.txt>.

Encoding

The standard specifies the information content of metadata for a set of digital geospatial data, it does not specify how this metadata should be encoded. The Clearinghouse however is proposing that the standard uses SGML to support metadata loading, exchange and presentation. A reference DTD is being developed to reflect the FGDC content element (draft versions are available from <URL:http://www.fgdc.gov/clearinghouse/dtds.html>). The General Record Syntax (GRS-1) will be used to encapsulate the metadata entry written in SGML.

Multilingual issues

There are no provisions made within the standard for the description/use of other languages.

Ability to represent relationships between objects

Within the implementation details (Annex D) there is a concept of a larger work citation that allows a data set to identify a parent metadata entry. These elements include:

- metadata entry level
- metadata identifier
- metadata series object identifier
- metadata data set object identifier

- metadata feature object identifier

These are internal elements that can be used to generate links between metadata objects. This model only supports ‘single inheritance’ e.g. a data set can only belong to one data series.

Fullness

Full - provides a very detailed content description for digital geospatial data sets with various specialised descriptive elements e.g. percentage cloud cover.

Protocol Issues

Protocol issues are not dealt with in the content standard but *Annex D* specifies TCP/IP and OSI protocols.

The Clearinghouse have implemented variants of the Z39.50 protocol, this was originally with the NISO/ANSI Z39.50 - 1988 standard using freeWAIS. The draft implementation document (*Annex D*) proposes the use of Z39.50 - 1995 compliant software and the GEO profile as a set of the FGDC metadata elements. The GEO profile is “is a draft specification for Z39.50 programmers that formalizes the data elements within the FGDC Metadata Standard as registered or well-known attributes in a Z39.50 server”(FGDC Newsletter, March 1996 <URL:http://www.fgdc.gov/News/fgdcnl0396.html>) , this will eventually allow queries by polygons and other spatial elements. The ISite ISearch software is being revised by CNIDR to support spatial search and to index SGML entries. There are also plans to support other database management systems such as mSQL.

Implementations

Ten sites are using the freeWAIS-sf software to implement the content standard. These sites are:

- UCSB Alexandria Digital Library (test subset)
- Montana State Library
- NASA Shuttle Earth Observation Photos
- NOAA Environmental Science Data Directory
- State of New York Clearinghouse
- Fish and Wildlife National Wetlands Inventory
- EROS Data Center
- USGS Water Resources Spatial Data
- USGS Geologic Information
- State of Wisconsin Clearinghouse

A prototype spatial data discovery system for Web clients is available through the National Geospatial Data Clearinghouse. This is providing a WWW gateway for the sites above who are using the freeWAIS implementation <URL:http://nsdi.usgs.gov/public/fgdcquery.html>. In addition the Clearinghouse with assistance from the US Geological Survey has developed some services in Brazil and Costa Rica and the Australian Environmental Resource Information Network (ERIN) has implemented a searchable Clearinghouse node for federal holdings in Australia.

GOVERNMENT INFORMATION LOCATOR SERVICE (GILS)

Environment of use

Documentation

The Government Information Locator Service (GILS) has a draft application profile (see <URL:http://www.usgs.gov/public/gils/prof_v2.html>) that is intended to be submitted to the Open Systems Environment Implementors Workshop/Special Interest Group on Library Applications (OIW/SIG-LA). It has already had an Implementors Agreement approved by the OIW in May, 1994. The Federal Information Processing Standard Publication (FIPS PUB) 192 (see <URL:http://www.ncsl.nist.gov/fips/fips192.wp>)

references the application profile. This specifies the GILS attribute set for Z39.50 for GILS servers and clients that support Z39.50 Version 2.

Constituency of use

The GILS was setup by the US Federal Government in order to provide the general public and its own employees with a means for locating useful information generated by the many government agencies. As such its constituency of use is very broad; literally anyone is likely to be able to search for resources using GILS and many different agencies are likely to use a variety of staff to generate their part of the overall GILS framework. Originally GILS was intended to force each agency to provide a set of locators that "together cover all of its information dissemination products" (Executive Office of the President, Office and Management and Budget, *OMB Bulletin*, no. 95-01, Dec. 7, 1994 <URL:http://www.usgs.gov/gils/omb95-01.html>). However in reality some agencies are using GILS as generic metadata records for many resources and others are hardly using it at all.

Ease of creation

GILS is a fairly complex metadata format, partly because of its breadth of coverage and partly because its design has been heavily influenced by the MARC and Z39.50 communities. Although it is possible that simple GILS records could be created by untrained staff, the format permits very rich and complex records to be created. As mappings to/from USMARC are provided in the GILS documentation it seems sensible to assume that at least some of the GILS records will be derived from USMARC records, which are themselves quite complex to create correctly.

Progress towards international standardisation

The GILS concept builds upon many international standards and has resulted in the creation of the GILS profile for Z39.50 servers and clients. This is gaining some support, mainly as a result of pressure from the US government, but is not nearly as widely implemented and deployed within the Z39.50 community as say BIB-1. Indeed, many of the targets linked to from the GILS information pages (see <URL:http://www.usgs.gov/public/gils/targets.html>) are just straight library catalogues with Z39.50 servers returning only MARC or SUTRS records and aren't even run by a part of the US Federal government. In the GILS community these servers are said to be providing "less than full GILS functionality." (<URL:ftp://ftp.cni.org/pub/forums/gils/log9604>)

Other comments

A number of other governments, such as the Canadian and Australian governments, are looking at the work done by the US GILS programme. Whether these are adopted on a large scale and what importance they will have in the future information society remains to be seen.

Format issues

Content

The information for this section is extracted from Annex E of the second draft of the GILS Application Profile. (*Application profile for the Government Information Locator Service*, Draft version 2. <URL:http://www.usgs.gov/gils/prof_v2.html>) This defines all of the elements in the GILS Core Element Set. These elements are defined as either being repeatable or not repeatable; the repeatable elements may appear more than once in a single GILS record whereas the not repeatable elements can only appear zero or one times.

Some of the elements are constructed from two or more subelements. For example, the *Controlled Subject Index* element is a grouping of subelements for *Subject Thesaurus* and *Subject Terms Controlled*. The grouping can be nested and is in this case; *Subject Terms Controlled* itself is a group formed from a repeatable subelement called *Controlled Term*.

Basic descriptive elements

The basic decriptive (bibliographic) elements included in the GILS Core Element set:

- Title
- Author
- Date of Publication
- Place of Publication
- Abstract
- Agency Program
- Resource Description

Subject description

The subject description elements included in the GILS Data Element set are:

- Controlled Subject Index
- Subject Thesaurus
- Subject Terms Controlled
- Subject Term Uncontrolled
- Controlled Term
- Local Subject Index

URIs

GILS Data Element records use the *Availability Linkage* and *Availability Linkage Type* data elements to specify the URI and MIME type respectively of the resource that the record is pointing at. These fields may be repeated within a single GILS record. There is also a set of *Cross Reference* elements that are used to refer to other, related GILS records. This set contains the *Cross Reference Linkage* and *Cross Reference Linkage Type* elements that indicate the URI and MIME type of the related record. The *Cross Reference* elements can also be repeated.

Resource format and technical characteristics

The resource format and technical characteristics and prerequisites are detailed in the following GILS Data Elements:

- Availability Medium
- Technical Prerequisites

Host administrative details

GILS Data Elements contain a number of fields to provide contact information. These include:

- Point of Contact
    - Contact Name
    - Contact Organization
    - Contact Street Address
    - Contact City
    - Contact State or Province
    - Contact Zip or Postal Code
    - Contact Country
    - Contact Network Address
    - Contact Hours of Service
    - Contact Telephone
    - Contact Fax

Administrative metadata

The administrative metadata required to maintain a GILS record is held in the following GILS Data Elements:

- Date of Last Modification
- Record Review Date
- Originator
- Control Identifier
- Original Control Identifier
- Record Source
- Schedule Number

Provenance/source

The GILS Core Data Element set provides the following elements for dealing with issues of provenance and record/resource data source.

- Purpose
- Availability
    - Distributor Name
    - Distributor Organization
    - Distributor Street Address
    - Distributor City
    - Distributor State or Province
    - Distributor Zip or Postal Code
    - Distributor Country
    - Distributor Network Address
    - Distributor Hours of Service
    - Distributor Telephone
    - Distributor Fax
- Sources of Data
- Record Source

Terms of availability/copyright

Terms of availability and legal restrictions on records and resources (including but not limited to copyright) are included in the following elements from the GILS Core Data Element set:

- Availability
    - Order Process
        - Order Information
        - Cost
        - Cost Information
    - Available Time Period
        - Available Time Structured
        - Available Time Textual
    - Access Constraints
        - General Access Constraints
        - Originator Dissemination Control
        - Security Classification Control
- Use Constraints

Rules for the construction of these elements

The definitions of the GILS Data Elements is given in Annex E (GILS Core Elements) of the GILS Application Profile. There is also a U.S. National Archives and Records Administration publication called "Guidelines for the Preparation of GILS Core Entries" (<URL:http://www.dtic.mil/gils/documents/naradoc/>). This is intended to specify which elements are mandatory in specific contexts within the US Federal Government, and also gives examples of customary usage of specific elements.

It should be noted that it is also permissible to use locally defined elements within GILS records in addition to the GILS Core Element set. Some of these elements may themselves be well known elements in other Z39.50 application profiles or other information systems.

Designation

GILS Data Elements are available in an extended attribute-value pair format and the GILS Application Profile also provides a mapping to and from USMARC Tags and GRS-1 record syntax.

Encoding

For physical transfer a GILS record may be delivered as a USMARC, GRS-1 or SUTRS record according to the GILS Application Profile. There may of course be further transfer encodings applied to these basic formats to allow the records to be sent through hostile environments. For SUTRS records, the GILS documentation defines an explicit preferred ordering to the output of the attributes so that the elements near to the top of the record are those most likely to show whether the record is useful to a searcher. However servers and clients are free to provide other orderings *in addition* to the preferred ordering.

Multi-lingual issues

The GILS Data Element set contains an element called *Language of Resource* that indicate the language of the resource that the record points at. There is also a *Language of Record* element that specifies the language that the GILS record itself is written in. Both of these are in the USMARC three character alpha code.

Ability to represent relationships between objects

The *Cross Reference* elements of the GILS Element Set provides for the ability to describe relationships between records. The *Cross Reference* element subsets are also intended to be used inside *Controlled Subject Index Subject Thesaurus* structures to describe where to acquire and reference the thesaurus.

Fullness

GILS is best described as being fairly *high* on the scale of fullness and complexity. For example in addition to all the elements described above for dealing with Document Like Objects (DLOs), it also contains a number of elements subsets for dealing with simple geospatial and temporal metadata. However it does not offer the range of specialised metadata formats that some of the more advanced geospatial applications require, such as percentage cloud cover.

Protocol issues

GILS servers are often implemented using Z39.50 servers, although this does not appear to be mandated absolutely as some US Government institutions are providing their GILS records by other means such as WAIS and HTTP. Some agencies are using the GILS records to generate HTML documents suitable for browsing by WWW browsers and some have either provided CGI front ends to their Z39.50 servers or loaded the records into another web accessible database in order to allow users to search their resources using a normal WWW browser.

Implementation

Most US Federal Government agencies now have GILS records deployed, and adoption of this format is being investigated by several other governments. Funding and encouragement from the US Government is also causing several companies, such as AOL and WAIS Inc, to start developing GILS compliant Z39.50 servers. Some of these will be freely available, whilst others will be commercial products.

IAFA/WHOIS++ TEMPLATES

Environment of use

Documentation

IAFA (Internet Anonymous FTP Archive) templates were designed by the IAFA working group of the IETF (Internet Engineering Taskforce) and guidelines were published in the form of an Internet draft in July 1995. (Peter Deutsch, Alan Emtage, Martijn Koster, M Stumpf. *Publishing information on the Internet with anonymous FTP*. Internet Draft. (working draft now expired). <URL:http://info.webcrawler.com/mak/projects/iafa/iafa.txt>). Template formats were drawn up for the various categories of information present on FTP archives: images, documents, sounds; services such as mailing lists and databases; as well as mailing list archives, usenet archives, datasets and software packages. Bunyip are now leading development of a whois++ White Pages directory system, Digger, which uses whois++ templates, a variation on the IAFA templates.

Constituency of use

Much of the driving force behind the development of the templates came from private companies, in particular from Bunyip as part of their development of Internet navigational tools and directory services; and from Martijn Koster at Nexor as a personal initiative. The aim of the IAFA template designers was to construct a record format which could be used by FTP archive administrators to describe the various resources available from their own archives.

IAFA templates were designed to facilitate effective access to FTP (file transfer protocol) archives by means of describing the contents and services available from the archive. Over the last few years many organisations wanting to allow access to their data, whether documents, datasets, images or software, have made them available as archives accessed by anonymous FTP. The IAFA template format has now been developed for use with the whois++ protocol, chiefly through the instigation of Bunyip who are developing directory service software conformant to this protocol.

The original intention was that each FTP site administrator would be responsible for ensuring that IAFA templates were available for each file on their archive. This information would be available for individuals visiting the archive and also, if FTP archive sites followed a common set of indexing and cataloguing guidelines, then it would be possible for software (such as Harvest) to automatically pick up the records. This is in fact happening in some implementations of the IAFA/whois++ templates, although in others records are being created

centrally. The recently developed directory service software, whois++, allows search and retrieval of databases created in this way, and also offers the possibility of searching across multiple databases. (RFC 1835 P. Deutsch, R. Schoultz, P. Faltstrom, C. Weider. *Architecture of the whois++ service*. <URL:ftp://ds.internic.net/rfc/rfc1835.txt>). Experimental work is being done using the Common Indexing protocol (CIP) which gathers together a 'centroid' or summary from a number of database to form an 'index server'. The index server contains an index of all unique attribute values contributed by the centroids, and searches can be referred from one index server to another by interlinking the servers in a mesh. (RFC 1914 P. Faltstrom, R. Schoultz, C. Weider. *How to interact with a whois++ mesh* IETF Proposed standard protocol, February 1996)

Supporters of IAFA templates have widened the original aim, and the intention now is to devise a record format simple enough to be generated by the wide variety of individuals and organisations involved with creating resources on the Internet, whether on web servers or FTP archives. The underlying philosophy is that it must be the information providers who create metadata records if indexing of the Internet is to be a viable proposition. Given the instability of network resources the alternative of centrally creating records would be a high cost option.

Ease of creation

The main advantage of the IAFA templates is that they are easy to create. IAFA templates are designed for use in a distributed system of record creation and storage so the simplicity of the records has been an underlying criteria in the design. Also they have been designed in relation to the objects they are trying to describe and are not hidebound by practices relating to non-electronic data.

Format issues

Designation

Resource types are identified by template type, and within each template type there are recommended attribute names to identify appropriate data elements. Template types which describe 'document like objects' (i.e.Document, Dataset, Mailing list archive, Usenet archive, Software package, Image, Video) all contain the same recommended attributes; other template types (Service, Mirror, Site configuration, Logical archives) contain their own specific attributes. The simplicity of the record structure is paramount, there is no allowance for identification of subfields, nor for 'qualifiers' to be attached to attributes.

Each record can only have one template type, but any of the other data elements can be repeated. It is intended that template types and data elements should be extensible, although extensions would not be inter-operable unless agreed between implementations.

Every time an individual or organisation occurs in a record there are a number of common data elements required to describe them e.g. name, address, telephone number, e-mail address. These logically grouped data elements are termed clusters in the guidelines and can be used to save indexing time by creating the details once then referring to them by a unique handle. The IAFA guidelines define the content of clusters for both individuals and organisations. Clusters of data elements can be identified by a unique handle although it is dependent on the implementation how the cluster information is incorporated into the record. Further proposals to extend the use of clusters have been circulated by Bunyip as part of the development of more detailed White Pages whois++ templates for use with the whois++ protocol. These proposals suggest definitions of further clusters at a lower level for names, phone numbers and addresses. In addition it is proposed that all clusters would include record management details.

Each record and cluster within the database is identified by a string of characters and/or digits unique within the system on which it resides.

Within the IAFA definition the repitition of attributes is achieved through the mechanism of variants. The first occurrence of an attribute is variant-1, the second variant-2. Related groups of attributes that are repeated are linked by the variant number e.g.

class-v1

class-scheme-v1

class-v2

class-scheme-v2

Within the whois++ schema the order in which attributes are stored is significant and links are maintained in this way.

Encoding

Records are held in simple ASCII text format. The syntax and semantics of data element names and values has been restricted to facilitate automated collection and indexing . Data elements are defined as attribute/value pairs and are of variable length. Attributes, record start and finish and continuation lines are recognised by the structure of the text and by insertion of defined 'special' characters. So for example continuation lines are signified by the first character being '+' or '-'; and records are delimited by blank lines.

Effort has been made to ensure the templates are 'human readable' which means less processing is required to make the data understandable. This helps to ensure there is a low entry cost to implement the templates. Attribute names are therefore written in full.

Content

Basic descriptive elements

The content is deliberately limited in detail in order to ensure the record is simple to create. The content includes simplified bibliographic fields (title, author, publisher, language). It is possible to distinguish personal and corporate authors by the choice of either the user or organization cluster.

The IAFA templates distinguish persons and organisations by appending the USER (person) cluster or the organization cluster to a particular element eg

- Author-(USER*)
- Publisher-(ORGANIZATION*)

Subject description

There is provision in all templates for a free text description of the resource to be included. In addition there is a keywords attribute for additional subject terms. Within ROADS usage of the templates further fields have been added to allow for subject classification and subject classification scheme to be added.

URIs

URLs are used for location of resources.

Resource format and technical characteristics

There are a number of different template types defined within the guidelines to describe the variety of network resources available:

- Document
- Dataset
- Mailing list archive
- Usenet archive
- Software package
- Image

Other template types are designed for use in the context of FTP archives to provide information about a particular FTP site:

- Site configuration information
- Logical archives configuration

- Service (e.g. on-line catalogues, information servers)
- Mirror (details of sites which mirror files including information on frequency of update from the source)

The configuration files would be relevant for the automatic collection of records, and in a broader context, the service template would be used to describe free-standing resources.

Templates for 'document like objects' include attributes for the size, format, character set and method of access. The guidelines set down that different versions of the same resource are described as variants. If a resource has ‘the same intellectual content’ it is taken to be the same resource regardless of language or text format (ASCII, Adobe, Postscript, etc.).

Host administrative details

The content of the record is designed to take advantage of the context in which the record will be used, so URL and e-mail links to authors and publishers are included.

Provenance/source

A source attribute can be used to describe details of the original form of the object.

Terms of availability/copyright

Templates for services include attributes for authentication, registration, charging policy, access policy, access times, and access policy. Templates for 'document like objects' include a copyright attribute.

Administrative metadata

The content includes detailed record management information including the date a record was created, the date for review as well as details of the creator of the record. This allows for the development of automated record maintenance procedures. It allows system administrators to keep track of rapidly changing resources and allows for quality checks to be carried out at regular intervals.

Rules for construction of these data elements

The guidelines acknowledge that the content of particular fields must be standardised to allow for effective indexing and retrieval. The following data elements have rules defined for the form of content as specified :

- e-mail addresses: RFC 822
- host names: RFC 1034
- host IP addresses: defined in guidelines
- numeric values: defined in guidelines
- dates/times: RFC 822 amended by RFC 1123
- telephone numbers: defined in guidelines
- latitude/longitude: defined in guidelines
- personal names: BibTex (see separate entry for BibTex)
- formats of resource: RFC 1521

The diverse locations of these rules, and the relative lack of detail compared to traditional cataloguing manuals, will inevitably lead to inconsistencies in practice. It remains to be seen whether the indexing and retrieval software can ameliorate the inconsistencies or whether 'simplified cataloguing rules' will need to be drawn up.

Multi-lingual issues

The IAFA guidelines state that text within the template is assumed to be in English using the standard ASCII character set although, using the whois++ protocol, it is possible to change character set within a template for a particular attribute pair by means of a system message. In the European context a more sophisticated means of character set negotiation is needed, but this could be overcome by having an agreed character set between particular clients and servers not subject to on-line negotiation.

Ability to represent relationships

There is no means of indicating a parent/child relation between documents (analytics), nor to link documents with ‘continued as’ or 'replaced by’. However these links could be provided by the keyword and subject descriptor searching and not be built into the record structure.

Protocol issues

IAFA/whois++ templates are associated with the whois++ directory service protocol (RFC 1835. P. Deutsch, R. Schoultz, P. Faltstrom and C. Weider. *Architecture of the whois++ service.* IETF, August 1995 <URL:ftp://ds.internic.net/rfc/rfc1835.txt>). This protocol fits closely with the IAFA template structure in that it passes attribute/value pairs and allows limits on search by template type, attribute, value or handle.

Progress towards international standardisation

The documentation for IAFA templates is in the form of an Internet Draft ‘Publishing Information on the Internet with Anonymous FTP’ (Peter Deutsch, Alan Emtage, Martijn Koster, M Stumpf. *Publishing information on the Internet with anonymous FTP*. Internet Draft. (working draft now expired). <URL:http://info.webcrawler.com/mak/projects/iafa/iafa.txt>). This document is a working draft which has no status as a standard, however it is a well developed exploration of a metadata record format specifically designed for Internet use. Both Bunyip and ROADS project workers are putting effort into a revised form to incorporate developments with the whois++ template. Convergence of the IAFA and whois++ template structures is likely as more implementations interoperate in a whois++ mesh (RFC 1914 P. Faltstrom, R. Schoultz & C. Weider. *How to interact with a whois++ mesh.* IETF Proposed standard protocol, February 1996). Implementation of services using the templates should also provide impetus to further development and modification of the template, and will also provide justification for progress along the standards track.

As yet there is no agreed mechanism for controlling amendments and additions to the template structure. Establishing a means to communicate and control changes to the templates would be an essential step in the move towards a standard. Until then the tendency is for attributes to proliferate and for the overall structure to remain unstable.

Implementations

There are now several implementation using IAFA/whois++ templates. The ALIWEB search system was the first to implement IAFA templates and it did so in the context for which they were originally designed. ALIWEB was set up as an experimental approach to providing access to FTP archives. Although ALIWEB was technically successful, the effort required to encourage FTP administrators to create records describing their archives could not be sustained and ALIWEB was integrated into the already established CUI W3 Catalog in order to encourage information providers. The future of ALIWEB remains uncertain but at present it is operational and is mirrored at various sites world-wide. ALIWEB is at <URL:http://web.nexor.co.uk/public/aliweb/aliweb.html>.

Bunyip are now leading development of a whois++ White Pages directory system. Within the eLib framework, so far three projects SOSIG (Social Science Information Gateway) <URL:http://sosig.ac.uk> and OMNI (Medical Information Gateway) <URL:http://omni.ac.uk> and ADAM (Art, Design, Architecture and Media) <URL:http//adam.ac.uk> are using the ROADS software. ROADS uses IAFA templates for description of resources, and the current release (version 1 in beta test Oct 1996) incorporates the whois++ protocol.

Within the UK there are also other implementations. The Internet Parallel Computing Archive (IPCA) at the University of Kent uses IAFA templates for a database containing information on parallel computing (David Beckett. *IAFA templates in use as Internet metadata.* <URL:http://www.w3.org/pub/Conferences/WWW4/Papers/52/>). At the University of Manchester, a volunteer effort NetEc provides a database of resources in economics using the IAFA template as the basis for the record structure <URL:http://cs6400.mcc.ac.uk/NetEc.html>.

Environment of Use

Constituency of use

The Inter-university Consortium for Political and Social Research (ICPSR) established a committee in May 1995 to develop a structured standard to describe social science data sets. The committee was a response to a perceived need amongst the social science archive community for an international codebook standard (a codebook generally contains information on the structure, contents, and layout of a datafile or data set).

Documentation

Information documenting the proposed SGML DTD (Documentation Type Definition) and content for the codebook standard can be found at <URL:http://www.lib.umich.edu/codebook.html>.

Ease of creation

The standard is still being formulated, the committee will be meeting in October 1996 to agree on a final draft with the intention that implementations will begin before the end of the year.

Progress toward international standardisation

The ICPSR is an international organisation with membership from 325 colleges and universities in North America and several hundred institutional members in Australia, Denmark, France, Germany, Great Britain, Hungary, Israel, the Netherlands, Norway, South Africa and Sweden. The codebook committee was established to be representative of all the archives and includes a representative from CESSDA (Council of European Social Science Data Archives), as well as representatives from Canada, Denmark, Norway and Germany. The elements for the codebook were chosen by reviewing a series of guidelines and standards in use by the social science survey, research, archive, and technical communities. The lists below include some of the materials that were examined:

Guidelines that prescribe what the codebook itself should contain (content standards):

- Roistacher: 1980, A Style Manual for Machine-Readable Data Files
- Geda: 1980, Data Preparation Manual (ICPSR)
- Collins, Patrick and Jane Powers, 1991, The preparation of data standards for machine-readable data.
- US Bureau of the Census, Statistical Research Division, Statistical Design and Methods Extension to Cultural and Demographic Data Metadata: CDDM draft standard 1995.
- Federal Geographic Data Committee content standards for digital geospatial metadata

Standards that define how to describe the study:

- Standard Study Description: developed by and for data archives, Council of European Social Science Data Archives.
- ICPSR Study Description "Template" Manual
- Essex Study Description outline (based on the Standard Study Description)

Standards that establish rules for producing records for cataloguing:

- MARC
- ISBD-CF: The International Standard Bibliographic Description for Computer Files
- GILS: Government Information Locator System
- ISO: International Standards Organization: ISO 690-2
- Dublin Core: OCLC/NCSA Metadata Workshop recommendations

Descriptions of codebook elements produced as a by-product of computerised interviewing software:

- Health and Welfare Canada
- Computer Assisted Survey Methods, University of CA, Berkeley

Standards that establish rules for tagging the contents of the codebook text:

- OSIRIS
- TEI: Text Encoding Initiative DTD for SGML
- EAD: Encoded Archive Description DTD for SGML

Other comments

The standard is still in the development phase but the indications are that the initiative has wide support amongst the social science data archives, the ICPSR also hope that data producers and granting agencies will adopt and support the standard.

Format Issues

Content

There are 5 main sections in the proposed structure:

- Codebook header
- Study description
- Data files description
- Record and variable description
- Other study-related materials

Each of the 5 main sections contain further sub-sections and elements.

Basic descriptive elements

The basic bibliographic elements of the data set are described in section 2 Study description under the sub-section Citation:

- Title statement of data set
    - title
    - subtitle
    - parallel title
    - common abbreviation
    - study number - producer
    - study number - archive

Subject description

The description of subject is dealt with in section 2 - Study description under the sub-section Study scope:

- Subject information
    - keywords
    - topic classification

URIs

None

Resource format and technical characteristics

The format of the data set is dealt with in section 3 Data files description:

- Type of file - text, numerical, graphic, program source, etc.

Host administrative details

These are provided for in section 2 - Study description under the sub-section Citation:

- Distributor statement for data set
    - documentation distributor
    - contact persons
    - depositor
    - date of deposit
    - date of distribution

Administrative metadata

All administrative information is provided in section 1 - Codebook header. Sub-sections here include:

- Title statement for documentation
- Responsibility statement for documentation
- Production statement for documentation
- Distributor statement for documentation
- Series statement for documentation
- Version statement for documentation
- Bibliographic citation of documentation

Provenance

The source of the data set is provided in section 2 Study description under sub-section Citation, elements include:

- Production statement for data set
    - producer
    - date of production
    - place of production

Terms of availability/copyright

This information is provided in Section 2 - Study description under sub-section Data access:

- Data set availability
    - original archive where study stored
    - collection note
    - extent of collection
    - completeness of study stored
    - number of files
- Data use statement
    - restrictions
    - access authority
    - citation requirement
    - disclaimer
    - analysis conditions
    - other reanalysis conditions note

Encoding

An SGML DTD has been proposed. Codebooks encoded into SGML could also be used for the production of data definition statements for use by statistical analysis software such as SAS or SPSS. There is also a proposal to produce a TEI compliant base tag set.

Multilingual issues

Details of language can be found in Section 2 Study description:

- Documentation statement
    - Language (s) of written materials

Ability to represent relationships between objects

There are fields for citing bibliographic information about and/or links to related materials and studies.

Fullness

Full - provides a very rich and comprehensive description of data sets.

Protocol Issues

There are no specified protocols associated with this format as yet but the committee are looking at the possibilities of using Z39.50.

Implementations

This is a proposed standard, the developers have applied the DTD to some sample codebooks but they are not in use as yet.

LDAP DATA INTERCHANGE FORMAT (LDIF)

Environment of use

Documentation

LDAP, the Lightweight Directory Access Protocol, is documented in RFC 1777. It is worth noting that LDAP started out life as a means of providing access to the X.500 Directory Service for machines which did not have the necessary power to run the full Directory Access Protocol over OSI lower layers. Over time LDAP has evolved to the point where it is now possible to use it as a stand-alone directory service protocol in its own right - i.e. with no use of X.500 behind the scenes.

The most recent revision of the LDAP protocol, which provides the necessary additional functionality, has yet to be published as an RFC - and is currently undergoing major modifications.

LDIF, the LDAP Data Interchange Format, is currently only documented in the LDAP software distribution from the University of Michigan <URL:http://www.umich.edu~rsug/>.

Constituency of use

LDAP and the associated LDAP Data Interchange Format are of interest to us because of their adoption by Netscape Communications Corporation, who in addition to developing a *Directory Server* product, have also said that they will integrate LDAP support with their popular Netscape Navigator World-Wide Web browser.

X.500 and LDAP are primarily used for White Pages type applications, such as searching for email and postal addresses, and telephone numbers. This should not be taken to mean that White Pages applications are their only use - X.500 was originally intended to support a wide variety of uses, including automatic routing of email and public key certificate distribution via the X.509 standard. In fact, White Pages is the only application of the technology which has been deployed to any extent on public data networks - and even then only in a trivial capacity by comparison with the near-universal adoption of the World-Wide Web suite.

If Netscape choose to integrate LDAP with their browser product in such a way as to facilitate its use for non-White Pages applications, this would mean that the vast majority of World-Wide Web users (most of whom are currently reckoned to be using various versions of the Netscape Navigator product) would have a search and retrieval protocol delivered directly to their desktop. This would be a major advance on the current situation,

where search and retrieval operations typically take place via third parties - often using dedicated WWW servers or CGI programs.

The key issue here is to do with the *schema* used to represent information within the LDAP based directory service. The schema is the set of attributes a particular object within the directory may inherit from whatever *object classes* the LDAP server is aware of. This is a similar notion to object oriented programming in languages such as C++ and Java. LDAP need not necessarily conform to the schema definitions which have been made for X.500, but this would be bad for interopability. In any case, the X.500 schema definitions are based on numeric *object identifiers*, whereas LDAP holds attribute names and schema information as plain text.

Be this as it may, flexible LDAP client support in Netscape Navigator and other similar products would require knowledge of other schemas in addition to straightforward White Pages information - perhaps including support for downloadable schemas. At the moment this is all supposition, however.

What is clear is that a schema suitable for document-like objects, such as World-Wide Web pages, does not exist at the moment. It could be expected that some time would elapse before consensus was reached on this, but the commercial interest may spur its development on on a unilateral basis. A suitable basis for experimentation might be the schema definitions used by Paul Barker and David Thomas for the ABDUX project though note that this work is primarily aimed at reproducing the OPAC environment via X.500 <URL:ftp://cs.ucl.ac.uk/abdux>.

Ease of creation

Here, and for the remainder of this section, we will focus on the use of LDIF as the "LDAP metadata format."

The LDAP protocol provides native support for object addition, modification, and deletion - so in theory LDAP based user interfaces could be used for all of the regular maintenance activities. This would facilitate the use of, for example, authority files.

The LDIF format provides a mechanism by which the LDAP server's database may be updated without using the LDAP protocol itself. Records may be created using any text editor, and no specialist training would be required except in cases such as (for example) where an attribute's value was to be restricted to words selected from a controlled vocabulary.

It should also be noted that LDIF is capable of handling binary data, which needs to be encoded according to the *base64* convention used within MIME. This may not be readily generated by hand!

Progress towards international standardisation

LDAP version 2 was recently advanced to the IETF standards track. This is not the version which implements stand-alone LDAP - that being version 3, which is still under development.

See also the notes below on URIs and internationalisation. Schema standardisation also would seem likely to occur within the auspices of the IETF, since this is where the majority of LDAP development has taken place.

Format issues

Content

Basic descriptive elements

None defined as yet.

Subject description

None defined as yet.

URIs

An Internet Draft proposing a formalism for inclusion of URLs and/or URIs has seen widespread adoption within the LDAP/X.500 developer community. This is being advanced through the IETF's *Access, Searching and Indexing of Directories* working group - ASID.

Note that the format which has been adopted optionally includes a plain text label associated with each URI, and that multiple URIs may be associated with a single object.

Resource format and technical characteristics

None defined as yet.

Host administrative details

None defined as yet.

Administrative metadata

None defined as yet, though some may be inherited from existing practice with X.500 - e.g. record modification times and ownership.

Provenance/source

None defined as yet.

Terms of availability

None defined as yet.

Rules for the construction of these elements

None defined as yet.

Designation

This takes the form of a series of attribute-value pairs, where the attribute and the value are separated by a colon and possibly whitespace - i.e. similar to mail and news headers, and whois++/IAFA templates. Each attribute-value pair appears on a line of its own.

Where an attribute has multiple values, these may be written as separate attribute-value pairs, on successive lines. Where an attribute-value pair would be too long to conveniently fit onto a single line onscreen, the value may be continued by starting a new line with a single space or tab character.

Multiple objects may be packaged into a single LDIF format file, with a blank line separating one object from another. Each object may be allocated a unique numeric ID to distinguish it from other objects.

Here is a sample LDIF object using the conventional Internet White Pages schema:

```
[12368]
dn: cn=martin hamilton, o=loughborough university, c=gb
cn: martin hamilton
o: loughborough university
c: gb
email: m.t.hamilton@lut.ac.uk
labeledURI: http://hpc.lut.ac.uk/~comth/
objectclass: person labeledURIObject
```

Encoding

The LDIF record must currently be encoded as plain vanilla ASCII text. As noted above, binary objects may be embedded, but only if they are converted to base 64 encoding first.

Multi-lingual issues

There is no support for multiple languages or character sets. This seems likely to appear in LDAP version 3.

Ability to represent relationships between objects

Potentially, complex relationships could be expressed. Aliases are well defined as part of the X.500 heritage - these can be used to provide persistent location independent names for objects. Other relationships are typically White Pages oriented, e.g. manager.

Fullness

None as yet. Has the potential to be anything from miminal to rich.

Protocol issues

LDAP runs over a TCP connection, and has dispensed with the extra OSI layers used by X.500 - though it does still use a cut down version of the Basic Encoding Rules to convert its data into on-the-wire format.

LDIF objects may be interchanged using many other protocols, since they are written as text. In particular, they may easily be used as MIME objects within mail, news and HTTP messages. Searching is an integral part of the LDAP protocol and X.500.

Implementations

The only widely deployed server implementation is the freely available one which was produced by the University of Michigan. Now that its principal developers have left to work for Netscape, its future is unclear. Note that the phrase *widely deployed* in this context means widely deployed amongst the small community of Internet sites which run directory services of one sort or another.

Numerous LDAP based gateways from other protocols and dedicated clients are distributed with the University of Michigan LDAP release. Some other development has taken place elsewhere, e.g. with the development of World-Wide Web and CCSO nameserver gateways.

MARC (GENERAL OVERVIEW)

Within this report we investigate selected MARC formats which are of interest in the context of the study. This entry provides some general context for other entries (USMARC, UNIMARC, UKMARC).

Environment of use

Constituency of use

MARC (Machine Readable Catalogue Format) originated in the 1960s as a means of exchanging library catalogue records. MARC was a response to the need for a standardised format for co-operating libraries to exchange and share catalogue records. It also met the requirements of national bibliographies for a format for their printed bibliographies, and it was used by bibliographic agencies for their supply of records to libraries. As library systems became computerised, MARC was used in library automation software as the basis for manipulating library records for display and indexing.

As use of MARC became more widespread the format was developed and adapted by its various user organisations and institutions according to their own disparate requirements. So, for instance, national libraries have tended to develop national MARC formats suitable for the material types they themselves catalogue (e.g. USMARC historically has included differing formats according to the requirements of the Library of Congress); library automation systems have developed MARC variants according to the needs of their users (e.g. the UK library management system vendors SLS and BLCMP use variants of UKMARC); adjustments for multilingual and cultural requirements have led to other formats (e.g. IBERMARC, CATMARC).

MARC format provides a means of integrating metadata into existing systems. National bibliographies, bibliographic record supply agencies, and individual libraries all have large collections of existing MARC records and want to integrate ‘Internet descriptions’ into their systems, and for them MARC is an obvious choice as it means their basic retrieval software can still be used to offer an integrated solution

The problem is that there is a considerable timelag between changes to the MARC format and changes appearing on the libray's OPAC, as this requires upgrades to the library's existing library management software.

Documentation

The proliferation of MARC formats was possible because the 'MARC standard' ISO 2709 *Format for bibliographic information interchange on magnetic tape* only governs record structure or encoding, it does not prescribe the content of the record within that structure. The record content of different MARC formats is defined in 'de facto' standards, usually controlled by national libraries; these take the form of cataloguing manuals outlining the formats and offering guidelines for their use.

Progress towards international standardisation

There is a move towards convergence of MARC formats: English speaking countries are converging to USMARC; European countries are looking to UNIMARC to map divergent MARC formats.

Convergence of other national MARC formats with USMARC mean that developments within USMARC are becoming increasingly influential. The critical mass of libraries using this format mean that it is possible for projects such as Intercat to find interested participants. The large numbers of USMARC users could be used as leverage with library automation system vendors for inclusion of format changes to be applied in OPACs, although this is an outstanding problem for Intercat participants.

Ease of creation

The creation of high quality MARC records requires training and experience in the use of cataloguing rules. Although the input of data to a record can be automated, the cataloguer needs to view the resource in some detail, and then interpret cataloguing rules before formatting the required information correctly in the MARC record. Within libraries specialist staff are assigned to cataloguing, they are often familiar with only one national format and indeed often specialise by material type (e.g. by subject area or format).

Format issues

Encoding

ISO 2709 states that a MARC record must consist of variable length fields with content designators; the record should have a record label, a directory, data field separators and record separators. Within these constraints different implementations of the standard have used different numbered tags and different subfield codes to identify the same type of bibliographic data.

The standard allows for optional indicators to appear after the tag and these are used in many MARC formats to qualify the tag. Within the text of a record, embedded subfield identifiers can be used to further identify data elements. ISO 2709 allows for fixed length data as part of the record label and this is used to store codes relating to material type, language, dates and so on.

Protocol issues

The Z39.50 protocol which enables search and retrieval of bibliographic information over the Internet is particularly designed to accommodate the search and retrieval of MARC records. The protocol can be used to pass searches of MARC fields from a Z39.50 client to a Z39.50 server fronting a databases of MARC records; and retrieved records can be returned in MARC format. The Z39.50 protocol uses attributes to identify how search terms should be treated by the server in a search. The bib-1 attribute set is defined in the standard and within that set the 'Use Attributes' were designed to map onto bibliographic records such as MARC. The bib-1 Use Attribute set does not contain any location or other non-bibliographic data so it is not possible to search on these fields. The protocol does allow for delivery of MARC records in full or abridged versions. There is no attempt in the standard to identify whether records searched or delivered are in the US or UK or other MARC formats. This can cause problems for interoperability in, for example, author personal name searches where the name is stored differently in US and UK MARC; similarly because the 'flavour' of MARC format is not identified it is not easy for the client to vary the display depending on the MARC format of the retrieved records. The standard allows for delivery of holdings information in OPAC records. At present the electronic address (and other non-bibliographic information) is not part of the bib-1 attribute set and is not searchable, however it would be displayed in a retrieved MARC record.

The majority of library automation systems allow for input and retrieval in MARC format, even if the records are stored internally in another format. Any changes in MARC, particularly regarding the display of the 856 field, will need to be reflected in OPAC software.

USMARC

Note: See also entry for MARC.

Environment of use

Documentation

The record structure of USMARC, as for all MARC formats, adheres to ISO 2709: 1981. This appears also as a national standard in the US, ANSI Z39.2.

The content of the USMARC record is not governed by an international standard, but by a cataloguing manual produced by the Library of Congress. The USMARC manual is issued by the Library of Congress and additions and amendments are controlled by the Library of Congress on the advice of the US MARC Advisory Group. This Group is made up of MARBI (The American Library Association's Machine-Readable Bibliographic Information Committee) and representatives from the US National Libraries, the National Library of Canada, the National Library of Australia, large bibliographic utilities (OCLC and RLIN), special library associations and library system vendors. The Library of Congress regularly publish discussion documents and proposals for comment which are considered at the twice yearly MARC Advisory Group meetings and, if agreed, are published occasionally as updates to the MARC format.

Constituency of use

Traditionally USMARC has been used by the library community in the US. Over recent years the national libraries of Canada and Australia have also adopted USMARC rather than maintaining separate formats. In addition the British Library have agreed on a timescale for a convergence programme with USMARC (see UKMARC section for details). Within the US the dominant bibliographic utility, OCLC, uses USMARC with some added variants. Widespread use of USMARC in the US enables sharing of cataloguing effort between the majority of academic and public libraries.

Over the last five years the USMARC community has been considering and adopting changes to the format to allow for cataloguing of electronic networked resources. At the start this was seen as an extension of the work done in the 1980s to describe computer files. The formats had become increasingly inadequate as a means of describing networked resources which needed to include details of access methods and addresses. After considerable debate within the USMARC community a new field, 856, has been adopted for the location of

electronic resources. Guidelines for its use have been issued by the Library of Congress. Changes for describing online services are still under consideration.

The Intercat project has been instrumental in progressing format development in this area. This is a project which started in July 1995 for cataloguing Internet resources and the timescale has recently been extended. It is led by OCLC with partial funding from the US Department of Education. The project involves participation of 200 libraries more than 60% of which are academic libraries; almost all active participants being in the US.

The Intercat project has catalogued a total of approximately 6000 resources to date. This relatively small number (as a comparison during a similar period OCLC's NetFirst database added 50,000 records, and webcrawler global search services many millions) reflects the co-operating libraries selection criteria. Libraries select only those resources they wish to integrate with their MARC library catalogues, and they select resources only if they are of sufficient quality and stability to warrant the effort of cataloguing. The project is by no means an attempt to 'catalogue the Internet'. There has been extended discussions on the Intercat mailing list on their criteria for selection of material. Many libraries intend to include in their catalogue only locally available resources i.e. those electronic resources held on a local server which are 'owned' or 'managed' by their own institution.

Progress towards international standardisation

The encoding of USMARC adheres to ISO 2709 in the same way as other MARC formats. The specific USMARC format is governed by a 'de facto' standard in the form of the USMARC manual produced by the Library of Congress. There is no move to formalise this as an international standard.

Format issues

Encoding

USMARC is an implementation of ISO 2709. USMARC records are written in the extended ASCII character set. The records consist of the leader, the directory and the data content fields. The leader consists of fixed fields containing coded data defining parameters for the processing of the record (such as the length of the directory entry), the directory contains entries listing the tag, starting location and length of each field in the record. The data content of the record is continued in fields of two types: variable control fields (fixed fields) and variable data fields.

Designation

USMARC formats are defined for three data types: bibliographic, holdings and authority records. This report will deal with the bibliographic record only. The USMARC Format for Bibliographic Data is designed for the description of different forms of bibliographic material: books, archives and manuscripts, computer files, maps, music, visual materials, serials. At one stage separate formats existed for each material type, but these have now been integrated.

Data in the record is contained in fields identified by a three digit tag. Fields containing data with a similar function is organised into groups identified by the first number in the tag:

- 0XX control numbers, provenance
- 1XX main entry
- 2XX titles and related information
- 3XX physical description
- 4XX series statements
- 5XX notes
- 6XX subject access
- 7XX added entries; linking fields
- 8XX series added entries
- 9XX reserved for local fields

The remaining numbers in the tags indicate further sub-division of content, and in general parallel content designation is preserved across the groups e.g.

- X00 personal author
- X10 corporate name
- X11 meeting name
- X30 uniform title

Further content designation is identified by a two character indicator following the tag, and by a two character sub-field markers. Within this scheme the digit 9 is used to indicate a local implementation. Most fields, and some sub-fields can be repeated.

The 0XX fields in the USMARC record contain fixed length data for information such as material type, date of publication, form, language.

Content

Basic descriptive elements

The MARC record is highly developed for bibliographic and bibliographic-like data.

USMARC developed in the context of library cataloguing. It therefore deals with the various bibliographic data elements in a detailed way. However it is important to note that the content of fields is governed by cataloguing rules. USMARC is designed to provide a formatted display of 'a catalogue card' giving a description of the resourceas well as to provide access for the purposes of information retrieval. The rules for the content within fields are governed by the Anglo American Cataloguing Rules, and the ISBD. So the content of the 1XX and 7XX tag ranges are defined in terms of the cataloguing concepts main and added entries and, depending on their relationship with the work, an author might appear in one or the other ranges. Similarly an author could be defined as a personal author, corporate author or meeting name and this will affect the indicator value.

In addition USMARC format has implications for the authority control of the data content. Normally the data in fields 1XX, 4XX, 6XX, 7XX, 8XX will be subject to authority control.

In order to integrate cataloguing of network resources into existing legacy USMARC databases it is necessary to create bibliographic data elements according to AACR2. As the Intercat project has discovered, this requires further extension of the cataloguing rules and extended guidelines if it is to be done in a standardised way.

Although library cataloguing data is still by far the most predominant use for USMARC, there are possibilities for use as a 'vehicle' for metadata created to other standards such as GILS or Dublin Core. MARBI Discussion Paper No. 88 raises some of the problems in such attempts, and in particular looks at the problem of defining a generic author field in USMARC.

Subject description

Specific 6XX tags are used for different controlled subject heading schemes e.g. Library of Congress subject headings, MeSH etc.

URIs

The 856 field has been approved for URIs. This field is designed to contain location and access information to make a connection, locate and retrieve an electronic document. Guidelines for the use of the 856 field have been issued by MARBI. This field may be repeated, and more than one access method may be used.

The 856 field is a structured field with subfields describing method of access, and it can be repeated to allow for different access methods. The 856 indicator details the mode of access over the network (e-mail, FTP, Telnet, dial up) or if none of these (e.g. HTTP, gopher, wais, prospero) then the mode can be defined in a subfield. Within the 856 field description of access methods, other than those taken from the indicator, follow the controlled vocabulary for Internet media types (also known as MIME types).

Resource format and technical characteristics

Non-bibliographic data is unstructured and tends to be placed in notes fields.

The MARBI Discussion Paper No. 49 presented a preliminary list of data elements to describe network information resources. This is developed in MARBI Discussion Paper No 54 (Providing Access to Online Information Resources). These papers map the required data elements onto USMARC fields and subfields.

For example there are proposals for:

- Type of resource — 256$a File characteristics
- Frequency of update — 310$a Current frequency
- Other providers of a database — 582$a Related Computer File Note

The Intercat project has also involved discussion on its mailing list of the use of various other fields in this context. For example:

- Detailed contents e.g. list of web links — 505 $a Contents note

Administrative metadata

- record review date — no
- record creation date — (in record label?)

Provenance/source

Not relevant.

Terms of availability/ copyright

There are proposals to use the following fields:

- Access restriction notes — 506 $a Restrictions on access note
- Mode of connection and resource address — 538 $a Technical Details note
- Host administrative details contact — 856$m

Ability to represent relationships between objects

USMARC uses tagged links to indicate relationships between parts of a collected work. Tags can also be used to specify other relationships e.g. continued as; replaced by.

Fullness

Allows for rich descriptions and detailed structure. See the entry for MARC.

Implementations

Widely implemented and deployed. USMARC also influences other MARC formats.

UKMARC

Note: See also the entry for MARC.

There are many similarities in the structure and content of USMARC and UKMARC. However the differences are significant and sufficient to make high quality conversion between the two formats complex. In addition the use of variants in the UK (e.g.. BLCMP and SLSMARC) and the minor differences in the US (USMARC and OCLCMARC) means high quality conversion between particular datasets need apply additional algorithms. UKMARC and USMARC are set to converge: the impact of this is not yet clear.

Given the planned changes to UKMARC, this review will give only a brief overview.

Environment of use

Documentation

In the UK the national standard appears as BS 4748: 1982. (*Specification for format for bibliographic information interchange on magnetic tape.* London: British Standards Institution). The *UKMARC Manual* (2nd ed. 1980) is published by the British Library and consists of a loose-leaf publication with several updates. The British Library National Bibliographic Service (NBS) is responsible for the UKMARC format. There are less complex procedures for agreeing amendments than for USMARC. The British Library (BL) introduced consultation procedures in 1992 whereby BL's proposals initially go for comment to the Book Industry Commission (BIC) Bibliographic Standards Working Party Technical Subgroup. The Subgroup is made up of UK representatives of the different library sectors, book suppliers, bibliographic utilities who are also library system vendors, as well as the NBS. This is followed by a period of public consultation with proposals included in the BL *Interface NBS Technical Bulletin*. After a period for comment the proposals may be adopted according to the final decision of the BL.

Constituency of use

Within the UK the majority of libraries have used UKMARC. In recent years some academic and national research libraries have moved to USMARC.

Comments

The 856 field has not yet been incorporated into UKMARC. At present there is an outstanding proposal from the British Library to adopt the USMARC 856 field as part of the convergence between UK and USMARC. This proposal will be considered in 1996 as part of the consultation procedure outlined above. Discussion of the detailed implications of convergence are now starting, particularly on the UKMARC e-mail list. As yet there has been little significant discussion regarding the more specific issue of cataloguing Internet resources using UKMARC.

Implementations

LINK, the emergent replacement to BUBL, and the associated Catriona project are planning to catalogue electronic resources using UKMARC. The few other UK libraries who are investigating cataloguing electronic resources tend to be using USMARC.

UNIMARC

Environment of use

Documentation

UNIMARC conforms to the ISO 2709 standard and to ISBD standards. UNIMARC: (the Universal MARC Format) was developed and published in 1977. The primary purpose of UNIMARC is to facilitate the international exchange of bibliographic data in machine-readable form between national bibliographic agencies, but in the UNIMARC Manual, first published in 1987, it was stated explicitly that UNIMARCs objectives would not only be conversion, but also a model for the development of new machine-readable bibliographic formats.

The latest edition of the UNIMARC Manual (2nd edition) was published in 1994.

A draft version of the UNIMARC Guideline 3 for Computer Files was issued in June 1995. These guidelines result from meetings of the IFLA Permanent UNIMARC Comittee and the requirements of the International Standard Bibliographic Description for Computer Files, ISBD(CF). A new draft of this Guideline is expected in July 1996. (The changes in the draft of the 2nd ed. of ISBD(CF), now being circulated for comment, will be applied to UNIMARC).

Constituency of use

The proliferation of national formats and the difficulty that resulted for the exchange of data was the main reason for the creation of an international MARC format which would accept, in principle, records created in any MARC format and act as a common format in terms of conversion. Since 1977 several national libraries have undertaken projects to convert from an existing national format to UNIMARC or have adapted UNIMARC for their national format needs. It covers monographs, serials, and cartographic materials, music, sound recordings, graphics, projected and video material, with provisional fields for computer files.

Ease of creation

UNIMARC was first developed as an exchange format and offers several options for description, so that records created on the basis of different cataloguing rules can all be included.

Progress towards international standardisation

The format is supervised by the Permanent UNIMARC Committee (PUC), under the auspices of the IFLA Universal Bibliographic Control and International MARC (UBCIM) Programme. Changes will be made only through the Permanent UNIMARC Committee. The content is described in the UNIMARC Manual (latest ed. 1994).

Although the PUC tries to maintain the standard, libraries implement the format in different ways, e.g. linking (4XX) can be used or not. In particular French libraries work with a variety of interpretations of the format.

Format issues

Designation

Data in the records is contained in fields identified by a three digit tag. Fields containing data with a similar funtion is organised into groups identified by the first number in the tag. UNIMARC consists of the following nine blocks:

- 0XX Identification block
- 1XX Coded information block
- 2XX Descriptive information block
- 3XX Notes block
- 4XX Linking entry block
- 5XX Related title block
- 6XX Subject analysis block
- 7XX Intellectual responsibility block
- 8XX International use block
- 9XX National use block

Content

Basic descriptive elements

UNIMARC deals with all the necessary bibliographic data. The following will concentrate on the adaptation of the format to enable input of data pertaining to online resources.

The Guideline 3 specifies the use of existing fields for the description of computer files, but in addition any other data elements from UNIMARC may be used in a record for a computer file. The probable need for additional fields or content designators and for redefinition of existing fields in the near future is acknowledged. Those should be brought to the attention of the IFLA UBCIM Programme Office.

Fields of which the use for computer files is specified include:

- Title (field 200): Title as it appears on container, box, opening screen, formal title screen, first display of information, header of the file etc.
- Parallel title (field 510): Title in another language appearing on the computer file.
- Author(s) (fields 200$f $g and/or 700, 701, 710, 711): Authors, programmers of the computer files as listed on the computer file.
- Author affiliation(s) (fields 700$p, 701$p, 710$p, 711$p): Institutional affiliations of the authors, programmers at the time the computer files were written or programmed.
- Edition statement (field 205): Any word or phrase indicating that the information was available previously in a different form.
- Publication, distribution (field 210).
- Physical description of the computer file (fields 215, 230): To be omitted for remotely accessed computer files, because there is no physical item.
- Accompanying materials (fields 215, 307): User handbooks.
- Series (fields 225, 410).
- Availability information (fields 345, 010, 011): Price units, stock number, agency for ordering a copy of the computer files.

Apart from the above mentioned fields, some of the (extra) information should be put in different fields of the Note block (3XX). This concerns the following data:

- Type of computer file (field 336)
- Technical details of computer file (field 337)
- Notes pertaining to title and statement of responsibility (field 304)
- Notes pertaining to edition: (Licensed by...) (field 305)
- Notes pertaining to publication, distribution (Shareware, etc.) (field 306)
- Notes pertaining to series (field 308)
- Notes pertaining to availability (field 310)
- Contents notes
- Users/Intended audience note

The 1XX block provides fields for:

- Coded data
- Qualifying data
- Language of computer file
- Target audience
- Publication date
- Country of publication or production
- Coded data relating to computer files: program, representational, textual.

Subject description

The 6XX Subject analysis block is used for subject data constructed according to various systems, both verbal and notational (e.g. UDC, DDC, Library of Congress Classification).

URIs

In Guideline 3 no special field is provided yet for information pertaining to location. USMARC 856 is being examined to see if it can be adopted for UNIMARC.

Resource format and technical characteristics

Field 135 is the provisional Coded Data Field for Computer Files. For type of computer file and technical details fields 336 and 337 in the Notes block are defined.

- In field 135, a one-character code indicates the type of data file:
  - a = numeric
  - b = computer program(s)
  - c = representational (pictorial or graphic information)
  - d = text
  - u = unknown
  - v = combination
  - z = other
- Type of computer file (field 336): contains information characterizing the type of computer file. In addition to a general descriptor (e.g. text, computer program, numeric), more specific information, such as the form or genre of textual material (e.g. biography, dictionaries, indexes) may be recorded in this field.
- Technical details note (field 337): This field is used to record technical information about a computer file, such as the presence or absence of certain kinds of codes or the physical characteristics of the file (e.g. recording densities, parity, or blocking factors). For software, data such as the software programming language, the number of source program statements, computer requirements (e.g. computer manufacturer and model, operating system, or memory requirements), and peripheral requirements (e.g. number of tape drives, number of disk or drum units, number of terminals, or other peripheral devices, support software or related equipment) can be recorded.

Host administrative details

No fields are specified for information pertaining to the host. USMARC practice may be adopted for UNIMARC.

Administrative metadata

There are no fields for record review date and creation date.

Provenance/source

Availability information is included in fields 345 (Acquisition information note), 010 (ISBN), 011 (ISSN). Further notes pertaining to availability go in field 310 (Notes pertaining to binding and availability).

Terms of availability/copyright

The relevant USMARC fields are being examined for this purpose.

Rules for the construction of these elements

Field 801 (Originating Source), subfield $g, contains an abbreviation for the cataloguing code used for bibliographic description and access. The Manual gives a list of the accepted codes in an appendix. Other codes may be registered with the IFLA UBCIM Programme.

Encoding

For data interchange in UNIMARC, ISO character set standards should be used.

Multi-lingual issues

Character positions 26-29 and 30-33 of field 100 subfield $a are used to designate the the default and additional graphic character sets used in the record. Sets approved for use with UNIMARC are:

- ISO 646 (IRV), Basic Latin set
- ISO 5426-1980, Extended Latin set
- ISO Registration #37, Basic Cyrillic set
- ISO DIS 5427, Extended Cyrillic set
- ISO 5428-1980, Greek set
- ISO 6438-1983, African coded character set

Ability to represent relationships between objects

The 4XX block is reserved for making tagged links to indicate relationship between objects.

Fullness

UNIMARC is a more concise version of the MARC format and compared with USMARC offers less richness of data, e.g. in description of materials.

The Guideline for Computer Files seem to have been formulated with offline products in mind ie. CD-ROMS, diskettes. No special fields such as URLs are specified for metadata specific for networked resources. In a table showing the data elements prescribed by ISBD(CF) and their corresponding UNIMARC locations, 'Access points: Technical details access' is referred to blocks 6XX and 7XX. The Guidelines are still being developed to be better suited for online materials as well.

Implementations

In *International Cataloguing and Bibliographic Control* (vol. 24, no 4, oct/dec 1995), a quarterly published by the IFLA/UBCIM programme, an overview is given of international UNIMARC Users and Experts. This list is the result of a questionnaire sent out in 1993 and updated in 1995. 35 of the total of 62 institutions that had replied indicated that they were currently using the UNIMARC format. Thirty of those institutions are located in European countries, the other five in the USA (Library of Congress), China (National Library), India (National Library), Japan (National Diet Library), South Africa (The State Library)

Central and East European countries especially seem to be interested in formats that guarantee easy access to international communities and in recent years there has been a growing interest in UNIMARC.

The need for easy exchange of information is recognized within the MARC communities. There is a programme for the harmonisation of the national MARC formats of Canada (CANMARC), the UK (UKMARC) and the United States (USMARC). Also, the British Library and the Library of Congress are committed to the development of UNIMARC. The Commission of the European Communities funded UseMARCON (User Controlled Generic MARC Convertor) project aims to develop a toolbox capable of converting bibliographic records from any MARC format into any other MARC format through a central conversion format.

PICA+

Environment of use

Documentation

Although the Pica+ format was in its design influenced by several MARC formats (INTERMARC, USMARC, UKMARC and UNIMARC) and follows ISBD standards, it doesn't conform to the ISO 2709 standard and therefore cannot be considered as a genuine MARC format (although internationally it sometimes seems to be considered as such).

The Pica+ format is not documented for external use. Cataloguers use the diagnostic format, which is a more user friendly presentation of the underlying Pica+ format. The diagnostic format is described in the *Richtlijnen voor de aanlevering van gegevens* (Rules for the input of data).

Constituency of use

Pica, the Dutch Centre for Library Automation, is a non-profit organization providing systems and services for the majority of Dutch academic and public libraries and for a number of library networks in Germany. Circa 200 Dutch libraries use their shared cataloguing system and about 400 libraries are connected to NCC/IBL, Pica's interlibrary loan and document delivery system.

Ease of creation

Pica is an extensive format, applied within libraries by specialist staff.

Progress towards international standardisation

Pica+ is not an international standard. For growing international cooperation and information exchange Pica has had to conform to exchange formats like UNIMARC and standard protocols like Z39.50.

Exchange between Pica+ and different MARC formats is possible (e.g. the USMARC records from RLG). Conversion programs for Pica+ to UNIMARC were made at the request of the German partners of Pica, and also to enable the STCN (Short Title Catalogue, Netherlands) database to be uploaded in the European database for publications from the handpress period that is being developed by CERL (Consortium of European Research Libraries).

Format issues

Designation

The Pica+ format has four digit tags (three numerals, followed by A-Z or @) and subfields. The diagnostic format has four digit numeric tags, the so called kmc's (kenmerkcodes or identification codes) and the subfields are marked with control signs that often correspond with ISBD punctuation. The Pica format has three levels, which in Pica+ (but not in the diagnostic format) are distinguished by the first digit, 0, 1 or 2:

- 0XXX General bibliographic level: contains all the fields that can be shared by all Pica users
- 1XXX Local level: the fields that can be used within one library or organisation. These fields are not visible to other users of the cataloguing system.
- 2XXX Copy level: fields to be used for one specific copy of an item. These fields are also invisible to other users.

The fields are divided in six groups, based on the kind of information. In the diagnostic format the tags of the fields within one group start mostly (but not all) with the same digit. The groups are:

1) Administrative data for the system and coded information

2) Discriminating data like ISBN

3) ISBD data (descriptive elements necessary to build an ISBD description, mainly used for export of titles, but also used for presentation)

4) Title and author index entries

5) Subject description: keywords and classification

6) Miscellaneous (administration, shelf number, acquisition etc.)

Content

In the E-doc project (KB, Pica and Surfnet BV) the Pica format used in the Shared Cataloguing System (Gemeenschappelijk Geautomatiseerd Catalogiseersysteem, GGC) was adapted to make the description and retrieval of online resources within the existing infrastructure possible. *Richtlijnen voor het catalogiseren van online resources* (Rules for cataloguing of online resources) were issued (Leiden 1995). These rules are a subset of the rules covering the description of audiovisual materials (*Richtlijnen voor de aanlevering van gegevens: audio-visueel materiaa*'). Only the rules that are different from the existing rules are included in this subset.

The format is still being tested, evaluated and adapted by a special Working Group: WG-FER, (Working Group Format for Electronic Resources).

New *Richtlijnen voor het catalogiseren van Computer Files* (Cataloguing rules for Computer Files, i.e. Online and Offline), which will replace the old guidelines for the cataloguing of - online and offline - resources, are being developed by this working group.

The free text general annotation field (4201) is being used for information about the (electronic) resource, for which there is still no field specified e.g. information about fees, passwords, subscription to discussion lists, login procedures etc. Possibly in the future the need for specified fields for some kinds of data will lead to further adaptation of the format.

Basic descriptive elements

The format is quite detailed in dealing with the various bibliographic data elements. The content of the fields is ruled by the Dutch interpretation of the ISBD standard.

Subject description

There are several fields (tags starting with 5 or 6) for subject description, on a general and local level. A shared system for subject description has been developed, consisting of a basic classification system (Nederlandse Basisclassificatie), and an additional keyword thesaurus (GTT).

URIs

There is one field (4083) for access (location en file) data of the online resource. The file format will be given, followed by one or more of the following subfields:

- =A URL
- =B file name
- =C path
- =D file size
- =E compression format
- =M connection type
- =N port number / protocol
- =O gopher type
- =P host computer name
- =Q host computer IP address
- =R name and location host organisation
- =S email address host organisation
- =T email address contact

File format and URL (subfield =A) are mandatory, the other subfields are optional. The whole field can be repeated.

Resource format and technical characteristics

- Field 4083 (see above).
- Field 4060: material type, e.g. text, image, video, audio, multimedia, software. If none of those apply, the type document' should be used.
- Field 4251 is specified for system requirements. This field is not yet specified in the above mentioned 'Richtlijnen', but is a result of recent WG-FER decision. Different subfields are specified for online and offline resources.
- A special field (4084) is specified for location and file data of images linked to the described document. Primary goal of this field is to present related images inside the same description. (If the described document has an image format, this will be noted in field 4083). In 4084 the file format will be mentioned, followed by one or more of the following subfields (only subfield =A is mandatory):
  - =A URL
  - =B file name
  - =C path
  - =D file size
  - =E compression type

Host administrative details

The official guidelines will be changed. Probably the new rule will be that fields for the publishers and 4030 and 4031 contain the geographical location and the name of the host, identified by the addition [host]. Still an item of discussion in WG-FER.

Administrative metadata

According to the official guidelines the annotation field (4201) is to used for record-last-verified (date of last check of availability of the online resource), and for record-last-update (date the description was changed for the last time), but this is also a point of discussion in WG-FER. The record-last-verified will be cancelled (every record has indications of creation and change dates anyway). The usefulness of record-last-changed is still being discussed.

Provenance/Source

Bibliographical links can be made from descriptions of printed to online versions and vice versa.

Terms of availability/copyright

In addition to the ISBN/ISSN fields, that contain also price information, the annotation field 4201 (free text) can be used for other data pertaining to availability and copyright.

There is no separate field for copyright statements.

Rules for the construction of these elements

The rules for the content within the fields are based on the *Regels voor de titelbeschrijving* (Cataloguing Rules), based on ISBD and created by the Federatie van Organisaties op het gebied van het Bibliotheek-, Informatie- en Dokumentatiewezen (FOBID). Pica adapted these rules for use within the Pica system.

Encoding

The character set used by Pica is a modified version of the INTERMARC character set, and is based partly on ISO standards 646, 8859-9 and 5426.

Multi-lingual issues

Field 1500 (mandatory) is a coded field for the language of the publication, the original text (in case of translations), the language of the abstract or the language of subtitles.

Ability to represent relationships between objects

There are several fields specified for the relation with other records or items. Those fields are not restricted to one special block of tags, like the 4XX block in UNIMARC. Links are made via the identification number of the record being linked to. There are two kinds of links: links to other bibliographic records (parts, supplements, other issues etc.) and links to records from authority files (e.g. the thesauri of author names, keywords).

Fullness

The format allows for rich descriptions, although only a number of fields is mandatory, so it is possible to stick to relatively simple descriptions.

Protocol issues

Pica uses its own proprietary session-based pica3 protocol for access to its own databases.

Implementations

Pica is in use by a great number of Dutch academic and public libraries. A comparable Pica system is in use by Die Deutsche Bibliothek and several of the German regional systems.

The format for description of online resources is being tested in a number of projects.

Environment of use

Documentation

RFC 1807 (A Format for Bibliographic Records by R. Lasher & D. Cohen. June1995) is a memo, not a standard, and defines a format for E-mailing bibliographic records of technical reports. RFC 1807 obsoletes RFC 1357 of July 1992.

Constituency of use

US technical community.

Ease of creation

The format was designed to be easy to read and create. Bibliographic records can be prepared and read using any text editor, without any special programs.

Progress towards international standardisation

RFC 1807 is not a standard in IETF terms, it is a memo.

Other comments

Programs have been written to map between RFC 1807 records and structured USMARC cataloguing records, and also from USMARC to this RFC.

Format issues

Designation

The format makes use of self-explaining alphabetic tags (field-ID's) to designate the fields. The four extra fields that were added to the old format (RFC 1357) are: handle, other_access, keyword and withdraw.

Content

Basic descriptive elements

- AUTHOR
- CORP-AUTHOR: Corporate author
- TITLE
- ORGANIZATION: publishing organization
- TYPE (e.g. summary, final project report, etc.)
- CONTACT: Contact for the author(s)
- DATE: Publication date
- PAGES: Total number of pages
- PERIOD: Time period covered (date range)
- SERIES: Series title, including volume number
- NOTES: Miscellaneous free text
    - Specific for technical reports:
- FUNDING: Name(s) of funding organization(s)
- CONTRACT: Contract number(s)
- GRANT: Grant number(s)

- MONITORING: Name(s) of monitoring organization(s)

Subject description

- KEYWORD: for controlled and uncontrolled keywords
- CR-CATEGORY: to be used for Computer Science publications. Possibly in future similar fields will be added for other domains. CR-CATEGORY contains The Computer Reviews Category according to the CR Classification System.
- ABSTRACT: not mandatory, but highly recommended. Unlimited in length, but applications should not be expected to handle more than c. 10,000 characters. The abstracts are used for subject searching.

URIs

Two fields are available for location and access data:

- HANDLE: Unique permanent identifiers that are used in the Handle Management System to retrieve location data. A handle is a printable string which when given to a handle server returns the location of the data. If the technical report is available in electronic form, the Handle must be supplied in the bibliographic record.
- OTHER-ACCESS: For URLs, URNs and other yet to be invented formatted retrieval systems. (Only one URL or URN per occurrence of the field).

Resource format and technical characteristics

No separate fields.

Host administrative details

There are ORGANIZATION and CONTACT fields for the publishing organization and the contact for the author(s) respectively. Their are no additional fields for data pertaining to the host organization providing the report, or the contact of the host.

Administrative metadata

- BIB-VERSION: identifies the version of the format used to create this bibliographic record.
- ID: identifies the bibliographic record
- ENTRY: Date when the bibliographic record was created
- END: Indicates the end of the record by repeating the same ID that was used in the ID field at the beginning of the record.
- REVISION indicates that the current bibliographic record is a revision of a previously issued record and is intended to replace it.
- WITHDRAW: indicates that the document is no longer available
- BIB-VERSION, ID, ENTRY and END must appear as the first, second, third and last fields and may not be repeated. The other fields may be repeated as needed.

Provenance/Source

- RETRIEVAL: Information on how to get a copy of the full text. Open ended format (= arbitrary text field)

Terms of availability/copyright

- COPYRIGHT: Copyright information of the cited report. Permissions and disclaimers. Open ended format (= arbitrary text).

Rules for the construction of data elements

The format is not designed for use with specific catoguing rules. Guidelines for the content of the fields are given in RFC 1807.

Multi-lingual issues

- LANGUAGE: The full English name of the language in which the report is written. If the language is not specified, English is assumed.

Ability to represent relationships between objects

Not many linking facilities. In the revision field a link is made to the obsolete record that is to be replaced.

Fullness

Medium. As the format is specially designed for the description of technical reports, a number of fields are only relevant to this kind of material and the format is not especially suited for the description of other kinds of documents.

Protocol issues

The format is designed for sending data of technical reports by e-mail. The RFC defines only the format of bibliographic records, not the way to process them.

Implementations

RFC 1807 is used by the Cornell University Dienst architecture (which provides an open, distributed digital library, of which all the services make use of the Dienst protocol) and by the Stanford University SIFT system (newsgroups).

RFC 1807 has been in use by the five ARPA-funded computer science institutions to exchange bibliographic records (Cornell, Stanford, UC, MIT and Carnegie Mellon University).

SUMMARY OBJECT INTERCHANGE FORMAT (SOIF)

Environment of use

Documentation

The Summary Object Interchange Format (SOIF) was designed as part of the Harvest Architecture developed at the University of Colorado at Boulder. It is documented in Appendix B of the Harvest User Manual <URL:http://harvest.transarc.com/afs/transarc.com/public/trg/Harvest/user-manual/node151.html>

Constituency of use

Records in SOIF are designed to be generated by Harvest gatherers and then used for user searches by Harvest brokers <URL:http://harvest.cs.colorado.edu/>. They provide a summary of the resources that a Harvest gatherer has found. The Harvest distribution contains a number of stock gatherer programs that can generate SOIF summaries from plain text, SGML (including HTML), PostScript, MIF and RTF formats.

In March 1996, Netscape Communications announced that they were also going to use SOIF in their catalog server product and a number of other search engine manufacturers are said to be looking at supporting it. Note that SOIF records could be generated by hand by archive maintainers or authors.

Ease of creation

The vast majority of SOIF records in use today are generated automatically by robots acting as Harvest gatherers. The format is a simple attribute-value based record and there is only a relatively small number of common SOIF attribute names, so it is easy to create SOIF records by hand if desired. As each Harvest broker can support any attributes that are required by the data it provides access to, it is possible for other attributes outside of the common set to be used in local systems. SOIF does not mandate any particular attributes, within the Harvest software it is possible to configure a customised record format (or template) which will be used within that particular implementation.

Progress towards international standardisation

SOIF is a really an internal record format of the Harvest and related systems and has not been placed on any formal standards track process. At the moment it is just a *de facto* standard.

Other comments

SOIF 'templates' are designed for a very specific purpose (summarising indexed resources) but the basic format is capable of being locally extended to handle other tasks if needed. However there does not appear to be any concept of nesting of elements in the SOIF format.

Format issues

Content

SOIF is based on simple attribute-value pair elements. A single SOIF stream can contain multiple SOIF 'templates', each of which has an URL for the resource that it refers to and a number of different elements for holding the other metadata. Each element has an attribute name, the length of the value in brackets, a colon delimiter and then the value itself.

Basic descriptive elements

The basic decriptive (biblographic) attributes in SOIF are:

- *Abstract*
- *Author*
- *Description*
- *Keyword*
- *Title*

Subject description

The common SOIF element set does not contain any subject description elements in the traditional library sense. It does however have a *Type* attribute name that describes what sort of resource the SOIF record refers to. The example types given in the Harvest User Manual are:

- Archive
- Audio
- Awk
- Backup
- Binary
- C
- CHeader
- Command
- Compressed
- CompressedTar
- Configuration
- Data
- Directory
- DotFile
- Dvi
- FAQ
- FYI

- Font
- FormattedText
- GDBM
- GNUCompressed
- GNUCompressedTar
- HTML
- Image
- Internet-Draft
- MacCompressed
- Mail
- Makefile
- ManPage
- Object
- OtherCode
- PCCompressed
- Patch
- Perl
- PostScript
- RCS
- README
- RFC
- SCCS
- ShellArchive
- Tar
- Tcl
- Tex
- Text
- Troff
- Uuencoded
- WaisSource

Most of these types are related to different types of computer file formats and languages, which reflects SOIF's intended use in indexing network accessible objects.

URIs

There is a URL at the top of every template that is the URL of the resource to which the SOIF record relates. There is also a *URL-References* attribute that can be used to hold any URL references that are present within HTML objects being summarised. Lastly, the contact information for the gatherer that generated the SOIF record is also provided in four additional atttributes. The example SOIF template mentioned in the Harvest User Manual also has separate *Site*, *File* and *Path* elements to allow replicated copies of the object to be located.

Resource format and technical characteristics

The *Type* attribute detailed above tells us something about the type of the resource being summaried. There is a *File-Size* attribute that tells us how many bytes are in the summaried object. SOIF also allows the actual object to be embedded within the template using the *Full-Text* attribute. The example SOIF template mentioned in the Harvest User Manual also includes a *Required* element that specifies hardware and software requirements,

Note that as the SOIF format includes the length of each value after the attribute name, it is possible to embed any binary object in the template if desired (although that means that it may not be possible to edit it by hand).

Host administrative details

The common SOIF element set provides no fields for this purpose. However the example SOIF template mentioned in the Harvest User Manual includes a *MaintEmail* element that is the email address of the maintainer of the object.

Administrative metadata

The common SOIF element set provides the following elements to contain information concerned with the administration of the template:

- *Gatherer-Host*
- *Gatherer-Name*
- *Gatherer-Port*
- *Gatherer-Version*
- *Last-Modification-Time* (of the object)
- *MD5* (checksum of the object)
- *Refresh-Rate* (how many seconds after the *Update-Time* before the SOIF template should be regenerated; default of 1 month)
- *Time-to-Live* (how many seconds after the *Update-Time* the SOIF template is still valid for; default 6 months)
- *Update-Time* (the time that this SOIF template was last updated; this is a required element and has no default)

Other SOIF elements that are not mentioned in the common element set but which are in day-to-day use for holding administrative metadata are:

- *CheckedEmail* (the email of the person who checked the SOIF template if hand generated)
- *EnteredBy* (the name of the person who entered the template)
- *Entered* (the date that the template was entered into the database by hand)

Provenance/source

SOIF's information on the source of the data is held in the four gatherer information attributes detailed above and also the URL of the resource that the template summarises. Some templates also include a *Version* element that gives the version of the resource that the SOIF template summarises.

Terms of availability/copyright

The common SOIF element set provides no fields for this purpose. However the example SOIF template mentioned in the Harvest User Manual has a *CopyPolicy* element that specifies the copyright and access policy of the resource.

Rules for the construction of these elements

The contents of the common SOIF element set are described in Appendix B.2 (List of common SOIF attribute names) in the Harvest User Manual <URL:http://harvest.transarc.com/afs/transarc.com/public/trg/Harvest/user-manual/node153>

Designation

SOIF records are in a simple attribute-value pair format. The length of the value is explicitly represented in each element, allowing binary objects to be embedded in the template.

Encoding

The physical transfer of the SOIF record between the gatherer and the broker (or the broker and another broker) in Harvest is often a simple byte stream containing the raw SOIF template.

Multi-lingual issues

There is no specific multi-lingual support in SOIF.

Ability to represent relationships between objects

The *URL-References* attribute allows links that are embedded in summaried HTML objects (and any other objects that can contain URLs such as VRML files) to be held separately in the template. There is no inter-template linking mechanism in the common SOIF element set.

Fullness

The SOIF common element set is a very simple and designed for a specific purpose (summarising gathered resources) and so they have a fairly *low* fullness.

Protocol issues

SOIF records can be carried over any transport protocol that supports a suitable application protocol.

Databases of SOIF records can be searched via a variety of mechanisms. The most common is a broker CGI script that can be accessed via a normal WWW browser. Other brokers use WAIS front ends. It would be possible to produce a Z39.50 front end, though it is not known if this has been done.

Implementations

The original implementation of SOIF was in the Harvest system, which is still freely available. Netscape Communications is using it in its Catalog Server product and other commerical indexing and search engine vendors are believed to be looking at supporting it.

TEXT ENCODING INITIATIVE (TEI) INDEPENDENT HEADERS

Environment of use

Documentation

The Text Encoding Initiative Guidelines were published in 1994 as a result of an international research project which started in 1987. The guidelines consist of a 1400 page manual available in print form or as an electronic document on the Internet.

Constituency of use

Burnard describes the goal of the TEI project as

> to define a set of generic guidelines for the representation of textual materials in electronic form, in such a way as to enable researchers in any discipline to interchange and re-use resources, independently of software, hardware, and application area. (Lou Burnard. *The Text Encoding Initiative Guidelines.* <URL:ftp://info.ox.ac.uk/pub/ota/TEI/doc/teij31.sgml>).

TEI is a joint project sponsored by three professional bodies: the Association for Computers and the Humanities, the Association for Computational Linguistics, and the Association for Literary and Linguistic Computing. The project was funded jointly from the US National Endowment for the Humanities and the European Union 3rd Framework Programme for Linguistic Research and Engineering. At present the project has two years more funding from the US for tutorial and dissemination work. The academic community in the US and Europe have been involved in the project forming a number of committees to consider different aspects of the encoding guidelines.

The TEI initiative aimed to reach agreement on encoding text across a range of disciplines. The TEI Guidelines, despite their origins in the humanities and linguistics were designed to form an extensible framework which could be used to describe all kinds of texts.

The TEI Guidelines specify that every TEI text must be preceded by a TEI header that describes the text. The header specification was formulated as part of the project by the Committee on Text Documentation comprising librarians and archivists from Europe and North America and the overall layout is grounded in a cataloguing tradition .

The TEI header can be used in different operational settings. Firstly it can exist as part of a conformant text. In this context the header might be created by the author or publisher as part of the original encoding; or it might be created during the TEI encoding of an existing document when it is used in a research or archival environment. Researchers can use the header in the process of textual analysis or, as is the case in a growing number of text archives, TEI headers are used as a means of bibliographic control.

The TEI Guidelines suggest that headers can be used in a second way by those libraries, research sites and indeed text archives who wish to build up databases of records referring to TEI encoded text held at remote sites. The Guidelines lay down a framework for 'independent headers', that is headers that can be stored separately from the text to which they refer. Independent headers are free-standing TEI headers which can be used in catalogues or databases to refer to a remote TEI encoded text.

A third possibility, not outlined in the Guidelines, is that independent headers could be used to describe networked resources which are not necessarily themselves TEI encoded. It is in this third context that independent headers could be described as metadata in the sense defined in this review. (It is assumed that metadata should be capable of describing any networked resource, not that there must be a necessary relation between the structure of the electronic data in the resource and the metadata format.)

Ease of creation

The level of difficulty in creating TEI headers depends on the amount of detailed information entered in the header, and the conformance of the content to external rules such as AACR2. If an independent header is to be created which contains the same content as a MARC record with the same adherence to cataloguing practice then the same level of skill would be required as for library cataloguing. If the header is to include details on encoding, profile and revision (see below) then this also requires detailed knowledge of the text. However the ethos of the TEI Guidelines is flexibility: the level of encoding detail can suit the requirements of the situation. Thus it would be possible for an author or ‘publisher’ of an electronic text to create a simple TEI header. This header could then be elaborated if required by an archive administrator.

Although the Guidelines recommend that TEI independent headers should be detailed, this recommendation is in the context of an archive. It would be possible for metadata records to be created using simplified content not in conformance to AACR2. Indeed the need for a simplified version of the full guidelines has been recognised. A subset comprising a ‘manageable selection’ of the full DTD has now been issued as TEI Lite (Lou Burnard. *What is TEI Lite?* <URL:ftp://info.ox.ac.uk:80/~archive/teij31/WHAT.html>). This subset includes the majority of the TEI core tag set and is designed to be sufficient to handle most texts to a reasonable level of detail. TEI Lite is in use by the Oxford Text Archive for the encoding of its own texts.

Progress towards international standardisation

TEI headers are conformant to the international SGML standard. SGML is specified in an international standard ISO 8879-1986.

Format issues

Encoding

The TEI Guidelines define textual features in terms of Standard Generalized Markup Language (SGML) elements and attributes, grouped into sets of tags. SGML aims to provide for mark-up of text in schemas which are hardware, software and application independent. SGML allows for a family of encoding schemes each with their own document type definition (DTD). TEI is a particular instance of a DTD; one that offers an extensible

framework consisting of a core set of features with a variety of optional additions. Within TEI it is possible to build a customised DTD, appropriate to the document being encoded, by declaration of tag sets being used. The independent header has its own auxiliary DTD set out in the Guidelines.

Designation

SGML provides a framework for defining data in terms of elements and attributes. In SGML schemes these terms have particular meanings different from usage in other metadata. An element is a textual unit such as a paragraph; within the header an element would be a unit such as a title or author. An attribute gives information about a particular occurrence of an element and would be structured as an attribute/value pair e.g. in the Profile Description there is a <textClass> element to identify the subject headings themselves, and the controlled vocabulary used is identified by an attribute <keywords scheme=LCSH>.

The various elements in TEI are grouped into tag sets:

- core sets: elements likely to be needed by all documents
- base sets: element sets appropriate for particular classed of document e.g. verse, prose, drama
- additional sets: elements appropriate for the specialised or detailed treatment of text in different classes of document.
- auxiliary sets: elements with specialised roles e.g. the independent header DTD

The tag sets are extensible to enable mark up of new sorts of material.

The TEI header forms one of the two core tag sets available by default to all TEI DTDs. Presence of the TEI header is mandatory in a TEI encoded text. The TEI header is made up of :

- File Description: the bibliographic characteristics of the document and its source
- Encoding Description: editorial decisions regarding treatment of the text and details of the editorial process as well as decisions on the treatment of blank lines, indents etc.
- Profile Description: additional non-bibliographic information giving the context in which the text was produced e.g. language, details of participants, subject classification
- Revision Description: details of updates, amendments to the text.

Within the header, elements may be indicated as being in free prose, or as being structured statements.

The independent header has the same structure as the TEI header but more guidelines on content. The independent header has more mandatory and recommended elements and the Guidelines recommend it should contain structured information rather than unstructured prose.

Content

Basic descriptive elements

The File Description is the only mandatory part of the header and it contains bibliographic description of the resource in the form of title, edition, publication and series statements. Within each element there is detailed bibliographic information e.g. the title statement includes information on intellectual responsibility specifying author, sponsor, funder, principal researcher, and other contributions. The form of the author however is not included i.e. as being personal, corporate or a meeting.

Within the File Description the title, publication and source are mandatory for all TEI headers, but several more elements are recommended for independent headers.

The file description contains detailed structured information drawing on standards and practice in the library cataloguing tradition and is modelled on library cataloguing standards.

Conversion to other formats

Within the Guidelines there is consideration of the conversion of TEI headers to USMARC records. The Guidelines include detailed suggestions for mapping particular TEI elements to USMARC tags, but acknowledge that human intervention would be required to create a quality MARC record. There is no attempt in TEI markup

to identify the author ‘main entry’, neither is the personal name format prescribed. Much of the non-bibliographic information would have no definitive resting place in MARC and would need to be moved to Notes fields.

In the independent header the usefulness of the profile, encoding and revision descriptions would be limited for analysis purposes unless the text was TEI encoded. Much of their usefulness depends on pointers in the electronic text to the header, relating information together.

Host administrative details, URIs

There is no provision for including location information, library call numbers or electronic addresses within the header. There is no consideration within the Guidelines of the description of services so there is no provision for host administrative details. However the flexible nature of TEI means that the tag sets could be extended to include this information.

Subject description

In the Profile Description there is a <textClass> element to identify the subject headings for a text. If a controlled vocabulary is used to identify the subject keywords then the scheme is identified by an attribute e.g. <keywords scheme=LCSH>; for classification numbers schemes are identified in a similar way e.g. <classcode scheme=DDC19>; if a user defined scheme is used this is identified by the <catRef> attribute.

Resource format and technical considerations

Any revision history of the resource itself can be included in the revision description. All changes of the machine readable data should be included in this part of the header.

Administrative metadata

There is no provision for data about the header itself to be included in the header.

Provenance/source

Information on the source of the electronic text should be included in the source description. The editorial declaration within the encoding description also allows for an explanation of editorial policy in the encoding of the text e.g. if spellings were corrected.

Rules for the construction of these elements

Where structured information is included in appropriate elements then the Guidelines give rules which follow AACR2 and ISBD. Those elements that are unstructured contain free text.

Multi-lingual issues

The profile description is used to specify languages used in the document.

Ability to express multi-lingual characters depends on the implementation of the TEI header. The guidelines do not specify any one particular character set , as with all SGML markup the guidelines are software and application independent.

Ability to represent relationships between objects

The source description allows for analytic references to be included, where an item is part of a larger collection. The type attribute can be used to distinguish the main title from subordinate, parallel or other titles.

Fullness

The role of the TEI independent header is so flexible that it can include large amounts of detail to enable analysis of text or it can be used in a simplified version to provide a known audience with bibliographic access to a collection of documents. This flexibility might well lead to difficulties if record creation occurs in a distributed

model as the level of tagging complexity, the richness of the record content, might vary considerably. It is desirable that all the headers in a particular database should have a comparable level of detail. Unless there is uniformity in the level of detail across the database, retrieval will suffer. This difficulty in controlling the level of detail would increase in a distributed environment and could lead to problems with interoperability and record sharing.

Protocol issues

Independent headers can be manipulated, searched and retrieved by any software that deals with SGML records e.g. Panorama, but as yet there is no provision within Internet search and retrieve protocols for TEI headers. Some research work is proposed to incorporate SGML DTDs into the experimental URCs.

Implementations

There are few implementations. The majority of present implementations are in humanities archives e.g. the Oxford Text Archive (<URL:http://sable.ox.ac.uk/ota/>) and the Electronic Text Center at the University of Virginia (<URL:http://www.lib.virginia.edu/etext/ETC.html>). Related European projects include EAGLES - Expert Advisory Group on Language Engineering Standards (<URL:http://coral.lili.uni-bielefeld.de/~gibbon/EAGLES/rwpaper/node5.html>) and Multext-East (<URL:http://nl.ijs.si/ME/>).

UNIFORM RESOURCE CHARACTERISTICS/CITATIONS (URCS)

Environment of use

Documentation

A number of proposals and counter-proposals for URC formats have been made, usually by posting them as IETF Internet Drafts e.g.

- An SGML-based URC service, by Ron Daniel and Terry Allen
- Trivial URC syntax: urc0, by Paul Hoffman and Ron Daniel

In addition, various other formats have been mooted at one time or another as potential candidates for URCs. Since there have been a number of proposals, and no one clear favourite, this document will consider general aspects of the URC work and proposals, rather than concentrate on one particular URC proposal.

Documentation on the Uniform Resource Identifiers work can be found on the World-Wide Web at :

- http://www.acl.lanl.gov/URI/
- http://www.gatech.edu/iiir/iiir.html

Constituency of use

It is important to note that there is (currently) no URC *per se*. The term URC has generally been used to identify:

- long term cataloguing information pertaining primarily to on-line resources
- a standardised means of associating so-called *metadata*, or describing information, with objects - not necessarily for cataloguing purposes
- information used as part of the process of resolving a Uniform Resource Name (URN) to a URL or URLs
- information used by applications when selecting a particular instance of a resource from a number of possibilities, not necessarily as part of a URN lookup.

URCs started off life as the responsibility of the Internet Engineering Task Force's Uniform Resource Identifiers working group, which was chartered to investigate both URCs and Uniform Resource Names (URNs) - persistent location independent naming. In an unusual step for the IETF, the URI group was disbanded due to what was felt to be a lack of progress.

At the time of writing, an effort was under way to form a new IETF working group specifically addressing URC issues, and with a more focussed remit than the old URI group. Specifically: the new group would focus on developing a common carrier architecture which could be used to package various resource description formats, rather than attempting to standardise upon one particular preferred format.

Ease of creation

Proposals have concentrated on formats which are readily created and understood by both humans and computer programs - typically encoded as plain text. It has been assumed that specialist training would not be required for human beings, with the URC format typically being no more complex than an HTML document or the headers of an email message.

Progress towards international standardisation

Arguably, none. Some experimental implementations have been developed, but none has been widely deployed. This is not a pre-requisite for Internet protocol standardisation, but it is rare for a protocol to be standardised before it has been widely deployed.

Other comments

Despite the interest in long term cataloguing type information, most of the URC proposals which have emerged over the years have not addressed this - choosing instead to deal with simple technically oriented information such as the object's Internet Media type. A notable exception to this trend is the URC proposal, which attempts to address many of these considerations using an SGML DTD drawn from the Dublin Core work.

Format issues

Content

Basic descriptive elements

Typically a small number of attributes designed to contain information intended for automatic processing, e.g. selection between multiple replicas of a resource, or indexing by a Web Crawler type application. Some basic bibliographic details may be present typically in a simplistic form e.g. it may be possible to indicate an object's *author*, but not whether this is an institutional/corporate author, or an individual.

Subject description

This has not received much consideration, except within the SGML URC proposal.

URIs

All of the proposals deal with URIs explicitly, though in some circumstances it may be acceptable to have a URC which does not contain any URIs - e.g. when the resource is not available on-line.

Resource format and technical characteristics

Information about the resource format is typically provided using an Internet Media type. Some proposals also include other technical information such as size in bytes and transfer encoding.

Host administrative details

Not a major concern.

Administrative metadata

This is typically not present, though it may be possible to deduce by other means - e.g. HTTP headers.

Provenance/source

Not a major concern.

Terms of availability

Not a major concern.

Rules for the construction of these elements

Not a major concern.

Designation

Typically this takes the form of either attribute-value pairs, in the style of mail/news headers or whois++/IAFA templates, or SGML Document Type Definitions.

For example, in the *trivial URC scenario* referred to above, a URC for the popular Z Shell package could be written as:

```
=====
  ftp://ftp.math.gatech.edu/pub/zsh
  The Z-shell, a command interpreter for many UNIX systems
  which is freely available to anyone with FTP access. Zsh is more
  powerful than every other common shell (sh, ksh, csh, tcsh and
  bash) put together.  The maintainer is Richard Coleman,
  zsh@math.gatech.edu
  =====
  ftp://ftp.sterling.com/zsh
  A mirror site in the US
  =====
  ftp://ftp.cenatls.cena.dgac.fr/pub/shells/zsh
  A mirror site in France
  =====
  ftp://mrrl.lut.ac.uk/zsh
  A mirror site in the UK
```

Note the use of equals signs "=" as delimiters between instance information, and that the only information provided, aside from the URL, for each instance is a textual descrption - and even this is optional. In the trivial URC proposal, the ==== delimiters could be augmented with an Internet Media Type (MIME type) to indicate when an object was available in multiple formats. By contrast, the SGML URC proposal referred to above provides mechanisms for specifying additional semantics in the URC:

```
<urc>

  <urn>urn:x-dns-2:shells.unix.computing.subjects.int:zsh</urn>

  <author>Coleman, Richard</author>
  <author type="email">zsh@math.gatech.edu</author>

  <title>The Z-shell</title>
```

```
<subject scheme="abstract">
A command interpreter for many UNIX systems
which is freely available to anyone with FTP access. Zsh is more
powerful than every other common shell (sh, ksh, csh, tcsh and
bash) put together.
</subject>

<instance>
<coverage>Canonical distribution site</coverage>
<url>ftp://ftp.math.gatech.edu/pub/zsh</url>
</instance>

<instance>
<coverage>A mirror site in the US</coverage>
<url>ftp://ftp.sterling.com/zsh</url>
</instance>

<instance>
<coverage>A mirror site in France</coverage>
<url>ftp://ftp.cenatls.cena.dgac.fr/pub/shells/zsh</url>
</instance>

<instance>
<coverage>A mirror site in the UK</coverage>
<url>ftp://mrrl.lut.ac.uk/zsh</url>
</instance>

</urc>
```

In this case, parsing the URC is much more difficult, but there is the reward of being able to express complex relationships between objects within the URC framework.

Encoding

Human readable plain text encodings have been the norm for URC proposals. It should also be noted that most proposals have not made a distinction between the information being represented and its encoding, and have made no provision for multiple encodings of the same information.

Multi-lingual issues

Language and character set variants of an object have been considered in some of the URC proposals. Only the whois++ based scenarios appear to go any way towards addressing these issues when they arise within the URC itself e.g. when the abstract associated with a document-like object is available in multiple language or character set variants.

Ability to represent relationships between objects

Most URC proposals have effectively codified a small number of well known relationships, e.g. between URN and URL(s), between an object and its creator, and so on.

Fullness

Variable from minimal to rich, depending on the proposal selected. Most proposals err on the side of caution and use a minimal set of attributes.

Protocol issues

Some URC scenarios have been allied to particular protocols, e.g. whois++ and HTTP. HTTP seems to be of primary interest as a means of transporting URCs, which is understandable given the popularity it currently enjoys. Some protocols would not be particularly suited to shifting URCs around - for example, SGML URCs would need to be specially *packaged* for transport over whois++, since the protocol is optimised for attribute-value pairs.

The most likely scenario for the proposed IETF URC group would seem to be to register a top level Internet Media type for URCs (and/or metadata formats in general), under which various metadata formats could be registered. This would provide the necessary convention within the MIME framework for metadata formats to be transported in not just the World-Wide Web (via HTTP), but also in MIME enabled mail and news software. A sample application of this approach would be to provide machine readable announcements of new software packages, Web sites, and so on. It would also neatly sidestep the arguments over preferred metadata formats which have prevented any real progress from being made on URCs in the past. It should be noted that although URC development has not been particularly rapid, the drive to introduce parental control on the material available via the Internet has led to the formation of a number of URC style efforts, typically using metadata embedded within HTML documents or the HTTP protocol. Perhaps the most notable example of this approach is the Platform Independent Content Selection (PICS) work sponsored by the World-Wide Web Consortium. Whilst PICS is oriented towards censorship, the format used is not limited to this application.

It has been suggested from time to time that URC implementations should be capable of supporting searching, e.g. so that the URC associated with a particular URL can be determined. whois++ would appear to be the most popular candidate for this search capability, though other protocols including Z39.50 and X.500 have been suggested. A cut down version of X.500, known as the Lightweight Directory Access Protocol (LDAP - see RFC 1777) has recently been adopted by Netscape Communications Corporation, for use in their *Directory Server* product. Whilst this appears to be primarily aimed at White Pages type applications, such as discovering email addresses, their stated aim is to incorporate support for LDAP into the Netscape Navigator World-Wide Web browser. Such browser support, if handled carefully, would effectively make LDAP the protocol of choice for the search and retrieval of URC type information. However, it remains to be seen whether LDAP will be supported in the sort of open ended way which is needed for these applications.

Implementations

A number of experimental implementations of the various URC schemes have been developed - the WWW pages referred to at the start of this section contain pointers to them.

WARWICK FRAMEWORK

Environment of use

Documentation

There is one detailed account of the Warwick Framework (Carl Lagoze, Clifford A. Lynch, Ron Daniel, *The Warwick Framework: a container architecture for aggregating sets of metadata*. July 12, 1996. <URL:http://cs-tr.cs.cornell.edu:80/Dienst/Repository/2.0/Body/ncstrl.cornell%2fTR96-1593/html>)

Constituency of use

The Warwick Framework is a container architecture for aggregating metadata objects for interchange. It was proposed at the second invitational workshop on Metadata, jointly organised by UKOLN and OCLC and held in Warwick University in April 1996.

The need arose from consideration of the Dublin Core. The Dublin Core addresses a particular aspect of the metadata problem: it is a simple resource description format. It could be extended in at least two ways. Firstly, it could be extended to accommodate elements which contain other types of metadata: terms and conditions, archival responsibility, administrative metadata and so on. Secondly, it could be extended to incorporate fuller resource description or to take account of specialist needs. However, other formats are also designed for resource description of different levels of fullness and within different communities. The IAFA document template is an example of one such format, USMARC another, the TEI header a third. It is undesirable either that there will be one single format for resource description or that a single format be indefinitely expanded to accommodate all future requirements. The need to retain a Dublin Core optimised for its target uses together with the need to exchange a variety of types of metadata led to the proposed Warwick Framework. This is a container architecture for the aggregation of metadata objects and their interchange. (Although the Dublin Core and the Warwick Framework are related by shared involvement one does not imply the other in any way.)

The Warwick Framework, then, is a proposed container architecture for the interchange of metadata packages. A package is a metadata object specialised for a particular purpose. A Dublin Core based record might be one package, a GILS record another, terms and conditions another, and so on. This architecture should be modular, to allow for differently typed metadata objects; extensible, to allow for new metadata types; distributed, to allow external metadata objects to be referenced; recursive, to allow metadata objects to be treated as 'information content' and have metadata objects associated with them.

Although there is wide agreement that this is a sensible approach, the Warwick Framework has not been implemented at the time of writing.

Format issues

Three approaches to the container architecture have been proposed: MIME, SGML, CORBA. There is also interest in using GRS-1 record syntax within Z39.50 as a container architecture.

DESIRE: Peer Review Report

Project Number:	RE 1004 (RE)	
Project Title:	DESIRE - Development of a European Service for Information on Research and Education	
Deliverable Number:	D3.2	
Deliverable Title:	Specification for resource description methods: a review of metadata: a survey of current resource description formats	
Review Method:	Report Reading	
Principal Reviewer:	Name	Dr. Stu Weibel
	Address	OCLC
	E-Mail	stu_weibel@oclc.org
	Telephone	
	Fax	
	Credentials	Senior Research Scientist, OCLC Office of Research
Other Reviewers:	(if relevant)	
Summary:	Relevant	5
	State-of-Art	5
	Meets Objectives	5
	Clarity	4
	Value to Users	5
Specific Criticisms	1	
	2	I have made notes in my review of areas where obvious changes in the state of the art exist and are not reflected in the report, and that this is a problem endemic to the area, and unavoidable. It is my belief that the report accurately reflected the state of the art as of its writing.
	3	
	4	The report is clear as it stands, but additional discussion of specific issues and minor reorganization of some of the information in it would improve its usefulness to those struggling to establish the context for the many extant metadata schemes (that is, most of us).
Developer Response:	1	*(developer's response given to general comments below)*
	2	
	3	
	4	

Peer Review Report general comments from Stu Weibel:

(Within this section Developer responses are italicised)

This report is the single most comprehensive survey of metadatastandards and issues that I am aware of. It is a very useful resource in a rapidly changing environment that is characterized by a proliferation of communities with different understandings of resource discovery, different requirements for description, different vocabularies for describing similar concepts, and legacy systems that mitigate against convergence of approach and process.

The report provides a comparative morphology, if you will, of many of the emerging formats in the metadata arena. It provides a valuable resource both for those immersed in the technology and for those who are entering the fray and seeking educational information to support their understanding.

The rapidly changing environment of network resource description makes it a particularly valuable resource, but also exposes the problem that anyone trying to enter the field will inevitably have. Many important issues have changed in the month since it was first written: there is a revised Dublin Core set, additional workshops have taken place, network software vendors deploy new systems, the importance of some protocols wanes as others wax (does anyone still use Gopher and WAIS, for example?).

There is no protection from this, but it raises the question of how to manage and document such change, and what the role any static report such as this can be a volatile context. Even as a snapshot in time of the metadata ecology, it is an important resource. My own hope is that this document will become the basis of a dynamically updated clearinghouse of metadata formats. This isn't an ultimate solution, but is certainly an important step along the way. Michael Day's collection of pointers to crosswalks is a very important contribution in this direction as well (and, I suspect, has arisen from precisely thisconcern about rapid obsolescence).

If I may wax philosophic for a moment, I think this work points out an interesting contrast between the thrusts of digital library work in the UK and the US. The US approach has focussed resources on a small number of large projects, the substance of which is more speculative and further out on the horizon. One may expect interesting and innovative results from these projects, but it is hard to see how their results will fit into the information services or educational computing environments of the next 5 years or so.

The UK eLib projects appear to be smaller, more diversified, and more practically focussed. This seems to me to have higher potential for return, certainly as far as the practical issues of resource discovery go. I applaud and encourage efforts such as these. They feed my optimism about the difficult business of fostering progress that is founded as much in sociological change as in scientific or technical advances.

Some minor niggles and nits:

In the section Metadata and its Uses, the discussion on coupling of the metadata record and its referent should be elaborated... this is a critical issue that is widely misunderstood, and the implications of different approaches should be discussed (embedded metadata versus closely coupled versus loosely coupled).

(This will be addressed as part of the ongoing work within the project on resource description. Experimental work is being undertaken at UKOLN to investigate holding metadata separately from its referent and this will be fed back into the working papers of the project.)

The figure at the top of page 8 should be made more clear. It identifies the general categories of the formats discussed and hence provides a framework for organizing the various individual schemes discussed subsequently, but it needs further formalization. What do the axes mean? Left to write presumably suggests simple to complex... is there any significance to the 3-D nature of the boxes? Should each of the three boxes have a label? How are the contents of each related? Much of this is explored in the text, but a diagram of this sort should stand alone (and have a title and caption, as well as clear labels).

(The figure has been revised into tabular format to aid clarity.)

The Dublin Core work has changed substantially from the representation here, as the authors will no doubt know, having participated actively in those changes. My presumption is that other formats might have experienced similar changes. This is not a criticism, but rather an observation in support of the problems identified earlier in this review.

(It is our intention to update and improve access to this work as part of further project work on resource description.)

Others of the formats described in this report are more stable. It might be useful to add in the first section a brief discussion (or table?) of the characteristics of the schemes, to summarize explicitly the classification of the schemes as stable, experimental, formally controlled, open or proprietary...? This information is identified in the discussions of the formats, but some of it might be usefully extracted and summarized in the introductory section on general principles. As a practical matter, many readers will not dig deeply into the meat of the format descriptions, wanting only a brief introduction such as the first section provides.

A summary of general characteristics in the front of the report might be useful (or perhaps even a middle section,

that offers a thumbnail comparison of the schemes in tabular format?

(We will consider elaborating the Typology Table, and/or including more tables to provide this information as part of further project work on resource description.)

Congratulations on a job well done, a service to the community that I hope will be sustained.

DESIRE: Peer Review Report

Project Number:	RE 1004 (RE)	
Project Title:	DESIRE - Development of a European Service for Information on Research and Education	
Deliverable Number:	D3.2	
Deliverable Title:	Specification for resource description methods: a review of metadata: a survey of current resource description formats	
Review Method:	Report Reading	
Principal Reviewer:	Name	Tony Gill
	Address	Surrey Institute of Art & Design, Farnham, Surrey GU9 7DS, UK
	E-Mail	tony@adam.ac.uk
	Telephone	+44 (0)1252 722441
	Fax	+44 (0)1252 712925
	Credentials	Programme Leader: ADAM & VADS. The Art, Design, Architecture & Media Information Gateway (ADAM) is an Access to Network Resources project of the Electronic Libraries programme.The Visual Arts Data Service will store curated visual arts resources and resource descriptions.
Summary:	Relevant	5 (1 = poor, 5 = excellent)
	State-of-Art	4
	Meets Objectives	4
	Clarity	3
	Value to Users	5
Specific Criticisms	1	Small number of unsubstantiated assertions made
	2	Small number of excessive generalisations made
	3	Some terminology used without adequate definition
	4	Terms associated with specific metadata formats used inappropriately
Developer Response:	1	*(developer's response given to general comments below)*
	2	
	3	
	4	

Peer review report: general comments from Tony Gill

(Within this section Developer responses are italicised)

The Survey document attempts to provide background information about the pertinent issues to consider when selecting a metadata format for implementation, and consistently structured outline descriptions of significant metadata standards initiatives to date.

The document is split into two main sections; Part I is a discursive overview of metadata and the general issues relating to the description of networked resources for a variety of purposes, whereas Part II provides a more structured directory-style description of the key metadata initiatives worldwide.

Part I

Part I provides a generally coherent and accurate summary of the issues, although it is somewhat terse in places, with certain passages assuming a high degree of prior knowledge on the part of the reader (see Clarity, below). There are also a small number of generalisations and unsubstantiated assertions that, whilst not necessarily disputed by this reviewer, possibly warranted more detailed discussion. For example:

> "It is unlikely that some monolithic metadata format will be universally used. This is for a number of more or less well known reasons." (page 5)

Some brief explanation of these reasons would be helpful.

(Explanation now included.)

> "Newer approaches based on manual descriptions have initially tended to focus on servers, and not describe particular information objects on those servers [..]. The subject information gateways fall into this category." (page 6)

The scope of the term 'subject information gateways' should be defined in this context before making this type of generalisation, since there undoubtedly exist subject-based information services that do not fall into the category as described.

(Definition now included.)

There is also some apparent inconsistency in the discussion of the three-band model for classifying metadata formats. For example, band one, a conceptual class of metadata format postulated in the review, is described as being "relatively unstructured data, typically automatically extracted from resources and indexed for searching." The apparent inconsistency is between the assertion that "the data has little explicit semantics and does not support searching by field.", and the statement that "there are moves to develop a shared format for exchange, perhaps based on SOIF." The inconsistency is that the Summary Object Interchange Format "is based on simple attribute-value pair elements", and should therefore support searching by field.

(Wording has been changed in the text to aid clarity. The reviewer has not perhaps taken into account the flexibility of SOIF which can be used for records with very little structure. In addition 'searching by field' implies some level of delineation of the semantic content of a record over and above the two or three attribut-value pairs that would be typical of a Band One record.)

The three-band model creates additional difficulties, since some of the other formats do not conform well to the defining characteristics of their class; for example, Alta Vista, a popular web crawler using a metadata format of the type described in band one, supports limited searching by field using HTML tags inherent in the resource itself. Similarly, Dublin Core records do not fully conform to the description of band two metadata formats, since they offer a relatively straightforward mechanism for describing relationships between objects. Overall, the three-band model appears to be somewhat artificial, and does not appear to add much value or clarity to the discussion.

(Any one format may not have all the characteristics of the band in which it is placed, and a note to this effect has been added to the text. In a number of discussions this grouping has proved beneficial in identifying the differences and similarities between formats.)

Taken as a whole, however, the Overview is an accurate, concise and useful introduction to the pertinent issues.

Part II

The use of a consistent structure across each entry in Part II, the review of metadata formats, enables comparisons between diverse metadata formats to be made, and the structure itself provides a sensible and clear description of each format in the context of the broader issues of resource description as outlined in Part I.

The descriptions of each metadata format generally provide a good synthesis between an analysis of the format, and discussion of the broader factors affecting the development of networked information discovery and retrieval initiatives. The *Implementations* section in particular is useful for ascertaining which formats are attracting interest from the influential web browser developer community.

Comments on individual sections are below:

CIMI

Caution should be taken when equating OSI (a framework for describing communication protocol layers) and TCP/IP (a family of communication protocols).

(Ambiguity now removed from text.)

The CHIO demonstrator requires the use of an SGML browser such as Panorama, in addition to a generic web browser, in order to view the SGML-encoded documents.

(Information added to text.)

Dublin Core

The Conversion to other formats section could be updated to include a reference to the DC/USMARC crosswalk exercise.

EAD

'Hand lists' (in the museum and archive sense at least) are not equivalent to detailed catalogues, but are more akin to inventory lists.

(Text changed.)

EEVL/EELS

Both the description of EELS and of EEVL talk of the absence of alternative formats for use by the engineering community, yet no cross-referencing between the EELS and EEVL is made.

(Cross-referencing now included.)

FGDC

Describing mSQL as a search engine is potentially misleading; it is in fact a freely-available relational database management system.

(The text has been amended.)

TEI Headers

The assertion that the inherent flexibility of the TEI Headers *"might well lead to difficulties"* could usefully be elaborated upon by examples of the type of difficulties that could be encountered as a result.

(The original comments on the implications for interoperability and distributed record creation have been elaborated.)

Clarity

The style of writing throughout both parts of the document is necessarily technical in nature, with acronyms and often-obscure references scattered liberally amongst the prose; since no guidelines about the intended audience for the document were supplied, it has been assumed throughout this review that the document is aimed at a reasonably technical audience with some prior knowledge of the issues pertaining to information retrieval in the network environment.

The multiple authorship of the document occasionally results in noticeable changes in the prose style from section to section. This has a marginal impact on the clarity of the document as a whole.

The liberal inclusion of URL's throughout, while slightly detrimental to the clarity of the document in paper form, allow it to be employed as a useful starting point for more in-depth study, and reflects it's dual role as both a traditional paper document and an (arguably more useful in view of the hyperlinking capability) electronic resource.

(HTML versions of the document have been made available as it has evolved.)

A more serious barrier to clarity is created by the occasional use of terminology associated with a particular metadata format to describe another format; the most common examples are the misleading use of the term *Template* to refer to records, a practice that has developed amongst the ROADS/IAFA/WHOIS++ community (pages 73, 75), and the phrase *Document-Like Objects* (pages 44, 45, 84), coined and only loosely defined by example in the Dublin Core initiative and not defined in the document under review.

(Different communities tend to use different terminology and this is certainly the case with metadata. For example templates, schemas, formats are used to refer to the 'format' of a record by different communities. The reviewer refers to the SOIF section where indeed the Harvest User manual does make use of the term 'template' to reference both format and record. Wherever possible ambiguity has been removed in the text, but there will inevitably be some borrowing of terminology amongst authors who come from different communities themselves.)

Technical slang is also occasionally employed, for example *vanilla ASCII* (page 55), *on-the-wire format* (page 55). These should not, however, present much hindrance to understanding for a technical reader.

(Where it proves useful and enlivens the style technical slang has been allowed to remain.)

It would also be helpful for the term *use chain* to be defined.

(The meaning of this term can be gleaned from the context. It is a phrase in current use in the field.)

A glossary of acronyms, and possibly some technical terminology, would greatly increase the clarity and potential audience of the document, should this be considered worthwhile.

(*We will consider adding a glossary as part of further project work on resource description*.)

A small number of typographical errors, listed as an appendix to this review, were spotted during the review process.

Conclusion

Documents of this nature are extremely difficult to compile and present clearly, since the requisite information, which must be collected from sources throughout the world, assimilated and reorganised, is almost immediately out of date in such a rapidly-evolving field.

Nonetheless, this Survey is a valuable and timely attempt to provide a coherent overview of the current state of the art of networked resource description, providing as it does a reasonably detailed and consistently structured account of the majority of the significant metadata initiatives taking place globally.

The Survey's usefulness is significantly enhanced by its publication as an electronic resource, allowing the user to carry out more in-depth research by following hyperlinks to detailed information about individual initiatives and formats.

This document is almost certainly the most comprehensive (and for the time being at least the most current) introduction to the diverse metadata formats currently in existence.