Kent Academic Repository

Full text document (pdf)

Citation for published version

de Souza Baptista, Claudio and Pinto, Francisco and Kemp, Zarine P. and Ryan, Nick S. (2000) MetaCRIS: Metadata for Research Digital Libraries. In: The Fifth European Conference on Current Research Information Systems. , Helsinki, Finland

DOI

Link to record in KAR

https://kar.kent.ac.uk/22027/

Document Version

UNSPECIFIED

Copyright & reuse
Content in the Kent Academic Repository is made available for research purposes. Unless otherwise stated all content is protected by copyright and in the absence of an open licence (eg Creative Commons), permissions for further reuse of content should be sought from the publisher, author or other copyright holder.

Versions of research
The version in the Kent Academic Repository may differ from the final published version.
Users are advised to check **http://kar.kent.ac.uk** for the status of the paper. **Users should always cite the published version of record.**

Enquiries
For any further enquiries regarding the licence status of this document, please contact:
researchsupport@kent.ac.uk

If you believe this document infringes copyright then please contact the KAR admin team with the take-down information provided at **http://kar.kent.ac.uk/contact.html**

KAR Kent Academic Repository

University of Kent

MetaCRIS: Metadata for Research Digital Libraries

Cláudio S. Baptista, Francisco Q. Pinto, Zarine Kemp and Nick Ryan
Computing Laboratory, University of Kent at Canterbury
Canterbury, Kent CT2 7NF UK
Tel.: +44 1227 764 000 FAX: +44 1227 762 811
{cdsb1|fqp1|zk|nsr}@ukc.ac.uk

Abstract

This paper describes a metadata model for Research Digital Libraries called MetaCRIS. MetaCRIS is a hierarchical metadata model that enables querying of libraries at different levels of abstraction. The model allows users to browse the underlying data sets and narrow the search space as they go deeper into the hierarchy. It encompasses three levels of abstraction. At the top level, information is mainly domain specific and less data dependent. At this level, there are catalogues describing the collections of different research subjects stored in the DL and ontologies, which are generic concepts applicable to specific domains. At the middle level, metadata about individual research projects are stored. At the bottom level, different multimedia data types are represented. Resources supported include reports, publications, references, images, video, audio and maps. At this level, not only metadata but also the functions are data type dependent. This model can be implemented using three approaches. The first one is to implement MetaCRIS in a digital library for Intranet access. The second one, extends the digital library by adding Web access. The third approach adds Z39.50 protocol capabilities for searching and retrieving data sets and links MetaCRIS with other Z39.50 MetaCRIS in order to interoperate. Finally, integrating these approaches might generate a CRIS Web Portal powered by a Z39.50 search engine.

1. Introduction

The high demand for multimedia data from different domains together with the acceptance and evolution of the Web as the global network infrastructure for exchanging information have pushed research into a new era of information systems [14]. In order to accomplish the requirements of this new era, Current Research Information Systems (CRIS) should integrate different data sources from different domains enabling interoperability, provide support for multimedia data and make the data sets available on the Web for wide dissemination and accessibility. One solution to integration is to use Digital Libraries (DL). DL organise different research information into collections. Ideally, these collections can be queried in the same way as we query traditional library catalogues and can provide support for multimedia data. The main benefit of DL is that not only metadata but also data themselves can be indexed and retrieved. In this paper a metadata conceptual model for CRIS DL is proposed based on the Common European Research Information Format – CERIF 2000 metadata model. Moreover, different architecture scenarios are discussed for implementation of distributed CRIS DL over the Internet.

The rest of this paper is organised as follows. Section 2 presents metadata concepts and highlights their role in CRIS. Section 3, discusses Digital Libraries, presenting the MetaCRIS conceptual model which states how CRIS could benefit from the former. Section 4 discusses architectural and implementation issues. Finally, section 5 concludes the paper.

2. Metadata

Research in the metadata field has attracted strong interest in the Internet era. Metadata are usually defined as data about data [10]. It is not a new concept; having been used for a long time in the database community for describing schemas, in the library community to catalogue the resources available, in distributed systems to describe protocols that should inter-communicate, and in many other applications. However, two main issues: *resource discovery* and *interoperability* have renewed interest in using metadata.

- **Resource discovery**: there is a large volume of information on the Internet, which needs to be indexed for subsequent efficient and effective retrieval. Current search engines, such as Alta Vista, Excite and Google[#] do not use semantic information about the Web resources. They index them based on keywords, which results in poor precision and recall [13]. For example, if one submits a search about *research projects*, a very large result set can be retrieved where the resources are semantically unrelated to what the user is looking for. In order to solve this problem, some metadata standards, such as Dublin Core [17], have been proposed to provide semantic information about the resources thereby imposing some order on the chaos.

- **Interoperability**: the heterogeneity of different systems and data sets makes the exchange of information difficult, the so-called *interoperability problem*. Metadata have been used for providing semantic integration of heterogeneous systems so that they can inter-communicate among themselves. This interoperability should be provided at different levels of abstraction, which can range from data-dependent to application domain-dependent metadata. Examples of the former are protocols used to access data, data formats, authorisation issues; whilst examples of the latter include descriptive metadata such as title, author, subject, abstract, and the use of ontologies that model the knowledge domain and enable interoperability at knowledge level [6].

Several metadata standards have been proposed for different application domain which include geospatial applications FGDC/CSDGM [3], library catalogues MARC [7], and museums CIMI [11]. For the Web, the Dublin Core standard has been widely accepted and the W3C XML and RDF are also being used [15, 16].

In the CRIS community, the relevance of metadata has been highlighted as the way forward, as suggested by Jeffery: “The future of CRIS is – in a word – ‘metadata’” pp.2, [8]. One of the main efforts have been made in order to consolidate the use of metadata in CRIS is the new version of the CERIF 2000, which is composed of three models: fullCRIS, data exchange and metadata models.

[#] http://www.altavista.com, http://www.excite.com, http://www.google.com.

3. Moving CRIS towards Digital Libraries

Digital Libraries (DL) have recently attracted a lot of research [9]. They are based on the traditional library model as there are resources and services available. The main difference is that the resources are mostly in electronic format and the services are software operations (e.g. functions) on these resources. In a DL there is no notion of reservation as the resources are not physically borrowed by users. Also, replication of the resources which is a very desirable characteristic of a traditional library (e.g. each library would like to have a copy of important books and journals), is not the case in DL as the resources can be reached from a distributed network of integrated DL. Another DL feature is that it can hold not only traditional alpha-numeric data such as those supported in CRIS catalogues, but any kind of multimedia data including: photographs, satellite images, video and audio clips, documents, hyper-documents, maps, graphics and the like. More importantly, these multimedia data can be indexed using their inherent semantics, which enable content-based queries for images and video, fuzzy queries in documents, spatial queries in maps and so forth. Moreover, security, billing model, and copyright rules are also issues considered in a DL design model.

3.1 DL and CRIS

DL seem to be a clear evolution of CRIS catalogues. The main advantage of using DL is that not only metadata for resource discovery is provided, which helps users to find where information about a particular research is stored, but also the *retrieving* of the information itself, in this case the findings of a particular research project in the form of papers, reports, images, video and audio. CRIS have been modelled and implemented as catalogues of information about R&D. One of examples of such implementations is the catalogue prototype ERGO (European Research Gateways On-line)[##].

In this paper, a metadata conceptual object-oriented model for DL, the so-called MetaCRIS, based on the CERIF 2000 metadata model is presented. Figure 1 presents this model using a subset of UML [4]. MetaCRIS has different levels of abstraction.

At the topmost level, there are *collections* and *ontologies*. The *collections* classify the research projects into categories so that searching and browsing at high level of details can be achieved. Examples of such collections could be: computer science, environmental science, biotechnology, geography, and the like. The *collection* attributes include: *identification, title, description, responsiblePerson, period, and number of projects.* The *ontology* class contains knowledge representation which includes thesauri, controlled vocabularies, and semantic relationships between concepts such as specialisation, generalisation, composition, synonym, hypernym and hyponym. CRIS knowledge representation is limited to CERIF and other classification schemes with translators, such as the multi-lingual Ortelius thesaurus.

At the middle level, metadata about *research projects, people and organisations* are held. The *ResearchProject class* contains the following attributes: *identification, title, responsiblePerson, abstract, themes, timetable, researchers, resources, organisation, funding organisation, budget, status and URL.* The thematic attribute is derived from terms represented in the

[##] http://www.cordis.lu/ergo/home.html

ontology class. A *research project* can be composed of many subprojects and it can contain several *hyper-documents*. The *Person* class contains attributes extracted from CERIF: *personId, family names, first name, other names, sex, honorific title, qualifications, nationalities and URL*. Finally, the *Organisation* class contains the following attributes: *organId, acronym, address, telephone, fax, e-mail and URL.*

At the bottom level, the multimedia data sets, which represent the results achieved by a given research project are maintained. They are modelled as a specialisation of the *hyper-document* class which holds metadata about each individual metadata type including: *title*, *author, description, date of publication, event*, *format, type, and price*. The *hyper-document* class uses the composite design pattern to specify that a document can be a composition of text, video, audio, or image or any combination thereof which results in a hypermedia document [5].

Indexing of multimedia data can be implemented using metadata so that images can use content-based retrieval and text can use fuzzy retrieval. The SQL-1999 standard, when fully implemented by the major database vendors will help to standardise these multimedia accesses[2].

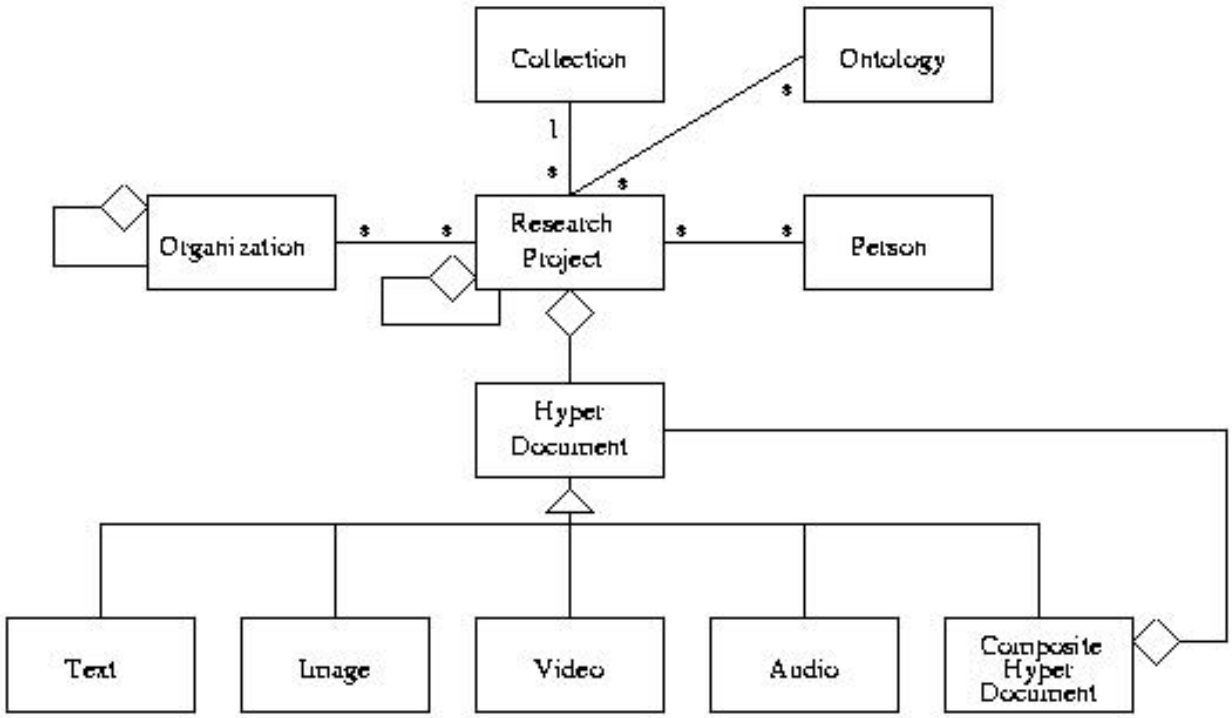

Figure 1: MetaCRIS conceptual schema.

4. Architecture

The MetaCRIS model can be implemented using at least three approaches. The first one is to implement MetaCRIS DL for an Intranet domain. The second one, extends the digital library by adding Web access. The third approach adds Z39.50 protocol capabilities for searching and retrieving data sets and links distributed MetaCRIS using Z39.50 as an inter-communication

technology with other Z39.50 MetaCRIS in order to interoperate. In the following, these approaches are discussed.

4.1. Intranet MetaCRIS

MetaCRIS is a DL supported by tasks such as storage, indexing, updating, searching and retrieving, giving the users access to the underlying data sets. A Database Management System (DBMS) is the natural application to support MetaCRIS (Figure 2), as it can provide persistence, query optimisation, indexing, concurrency, recovering, security, and other capabilities [12].

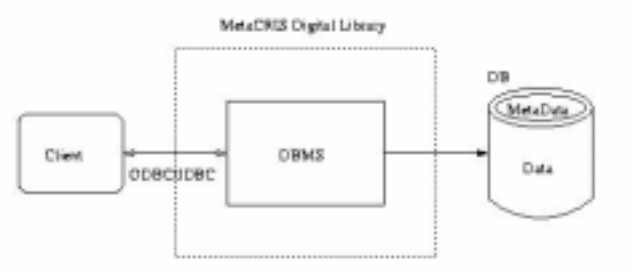

Figure 2: Intranet MetaCRIS.

Therefore, the basic implementation of MetaCRIS can be completely based on a DBMS accessible with private mechanisms such as SQL and closed in its context for an Intranet. However, other features are desirable for the distributed system, which can increase the system functionality and its scope in terms of visibility and interoperability.

4.2 MetaCRIS Web Access

Web Access is one of the most important requirements as it can offer the services provided by the MetaCRIS to a broad number of users. Web Access can be achieved having a Web Server as a DL front-end. Besides, to support users browsing, it should provide users with the functionality to search and retrieve all objects stored in the DL using a simple Web browser.

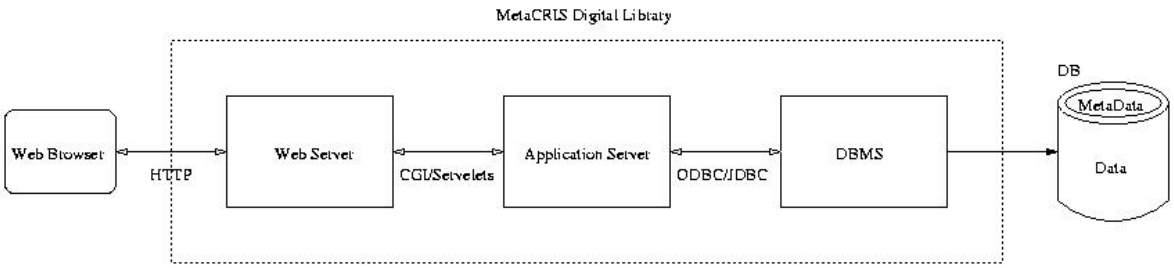

Figure 3: MetaCRIS Web Access.

Although with Web access it is possible to access the DL objects managed by the DBMS, these objects are not located in the Web space. The Web access is generated on the fly creating temporary links to a subset of the database objects in the Web Space for the objects matching a special search query. Moreover, even though MetaCRIS is now accessible from

the Web and other CRIS can be implemented with Web Access, nothing has so far been said regarding how distributed MetaCRIS can interoperate.

4.3 Distributed MetaCRIS: interoperable Z39.50 Access

In order to promote interoperability among distributed MetaCRIS over the Internet a possible approach could be adding to each MetaCRIS a protocol such as Z39.50 to access the underlying data sets.

Z39.50 is a protocol which defines data structures and interchange procedures allowing a Z39.50 client to specify queries with pre-defined terms (attribute set) to search databases on a Z39.50 server[1]. The servers abstract the databases, showing only a plain list of access points, making no assumptions of how the databases are organised internally.

Broadcast searching is not defined in the protocol, but it can be achieved by a special application that permits a client to search multiple Z39.50 servers in parallel. This application is built on top of Z39.50, and can distribute connections to heterogeneous machines using the Z39.50 protocol. Such an application can be a Gateway Web/Z39.50 located at the MetaCRIS Application Server, which receives requests from the Web, provides broadcast Z39.50 searching, collects the result sets and sends them back to the Web. Therefore, MetaCRIS users can access objects stored in the database, Web and Z39.50 spaces.

MetaCRIS can also interoperate with legacy CRIS, provided the latter implement a Z39.50 front-end with common access points. A Cross Domain (XD) attribute set is a common denominator to provide the semantic searching access points. As CRIS and MetaCRIS uses CERIF, therefore it is feasible to map them to a XD attribute set. This interoperable and distributed approach is an integrated basic system which generates transparent access to both the Web and Z39.50 spaces. This approach offers its own basic search and retrieval services in terms of semantic searching access points for the DL and can reduce storage and bandwidth requirements.

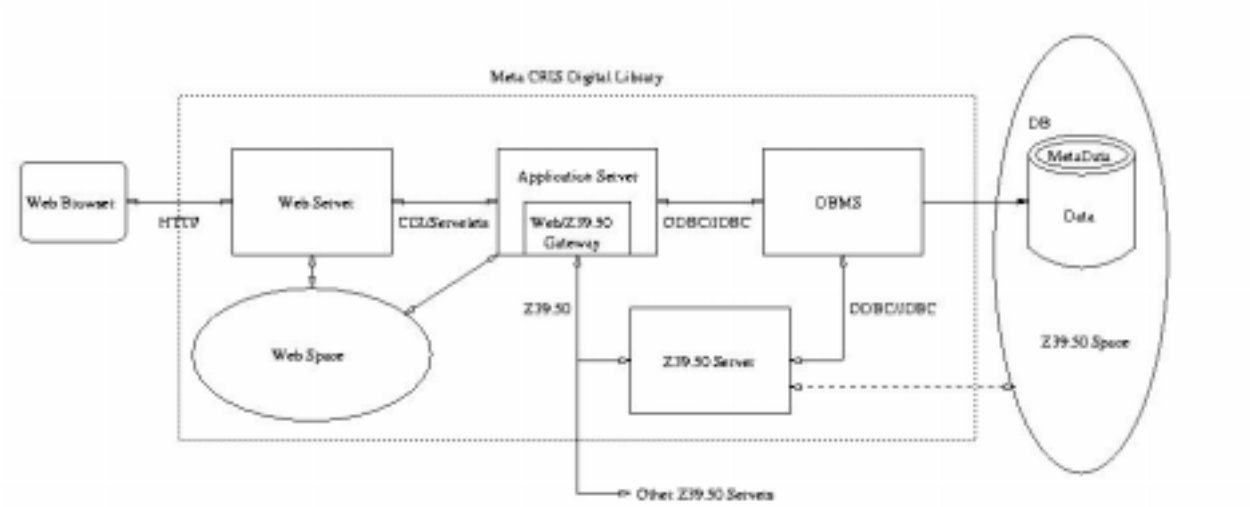

Figure 4: Distributed MetaCRIS: interoperable Z39.50 Access.

All the internal tasks of the DL continue to be done by the DBMS (e.g. searching, indexing, storage). A Z39.50 search and retrieval engine can help users to browse the subjects or choose the keywords for searching, provided that they have access to the Web. This is done using the power of Z39.50 to find and obtain the wanted objects in the original places and putting temporary database object links in the Web space.

4.3.1 Brief Explanation: a MetaCRIS walkthrough

As a starting point for the CRIS service, there is a Web Portal, which explains to the user the service proposal and shows some illustrated examples of how to use the service. If the user wants to specify which subject(s) to search, an assisted mode service is offered which drives the user to the required subject (collection), otherwise a Cross Domain query can be executed over all collections in any subject. If too many hits are received, the assisted mode service is offered again. This Web page is dynamically generated, based on a Web/Z39.50 Gateway, which automatically connects to a Z39.50 Server and obtains the information necessary to show the collection of research projects maintained in the DL.

The ontologies are accessible dynamically via the Web and are represented as trees. An application located in the Application Server connects to the database and generates automatically all ontologies in XML, allowing any XML capable browser to show the ontology tree. These ontologies are represented as XML files and are linked to each other by X-Links.

Having chosen the collection, the user can browse/search the research projects and eventually their sub-projects. For each research project, information about related people and organisation can also be accessed. Finally, having retrieved the research project metadata records, the user can retrieve multimedia objects related to them.

5. Conclusions

MetaCRIS is not a metadata catalogue but a full DL, which encompasses the metadata catalogue functionality and additionally provides access to the underlying data sets. MetaCRIS assists the users to navigate the DL at different levels of abstraction. The positive aspects of using MetaCRIS include: n*arrowing the search space*; *semantic metadata description; distributed and Internet accessible data and metadata; interoperability with other MetaCRIS; multimedia objects, indexing and retrieval; and multi-modal interfaces.*

MetaCRIS can use an application (robot like) to synchronise the metadata from the database to the Web space having a metadata surrogate copy of every meta object created in the database. However, current popular search engines such as Alta Vista do not take advantage of this. As they do not use metadata to index the Web pages - the so-called *metadata spam* problem*,* neither CERIF nor even DC will be seen by the Web robots.

A possible solution might be to develop meta-search engines, which use the metadata tags present in the CRIS DLs. This makes MetaCRIS metadata accessible to the whole Web space, having all the database objects visible from the Web space as if they were part of the Web. However, this approach acts as a mirror of the database, which raises some problems such as replication, consistency, security and storage.

From a global Web perspective it would be very difficult or even impossible to have an efficient and effective search engine based on robots or a directory system based on manual work. A DL should provide a complete range of services not only for the users but also for the applications such as meta-searchers which can take full advantage of these local services to create a efficient and effective distributed search service. If standards such as DC are adopted for the Web, search engines and directory systems can effectively access the semantic search points to the service. Therefore, distributed searching on the Web can be implemented feasibly without too much additional effort.

Acknowledgements

Cláudio S. Baptista and Francisco Q. Pinto would like to thank CAPES/Brazil and FCT/Portugal respectively for partially funding this research.

References

1. ANSI/NISO, Z39.50-1995 Version 3, Information Retrieval: Application Service Definition and Protocol Specification, 1995. Available at URL < http://lcweb.loc.gov/z3950/agency/> [referenced 10.03.2000].
2. Eisenberg, A. and Melton, J., SQL:1999, formerly known as SQL3, SIGMOD Record, 28(1), March 1999.
3. Federal Geographic Data Committee, Content Standards for Digital Geospatial Metadata, Workbook Version 1.0, National Spatial Data Infrastructure, Washington, D.C., USA, March 1995.
4. Fowler, M. and Scott, K., UML Distilled: Applying the Standard Modelling Language, Addison-Wesley, 1997.
5. Gamma, E. et al., Design pattern elements of reusable object-oriented software, Addison-Wesley, 1995.
6. Guarino, N. Formal Ontology and Information systems, In: Proceedings of the International Conference on Formal Ontologies in Information Systems (FOIS'98), Trento, Italy, June 1998.
7. ISO, ISO 2709:1996 Information and Documentation – Format for Information Exchange, 1996. Available at URL <http://www.iso.ch/cate/d7675.html> [referenced 08.12.1998].
8. Jeffery, K. The Future of CRIS, In: Proceedings of the CRIS'98 Conference, Luxembourg, March 1998. Available at URL <http://www.cordis.lu/cris98/> [referenced 10.03.2000].
9. Lesk, M. Practical Digital Libraries: Books, Bytes, and Bucks, Morgan Kaufmann Publishers, 1997.
10. Miller, P. Metadata for the Masses, UKOLN, 1997. Available at URL <http://www.ariadne.ac.uk/issue5/metadata-masses/> [referenced 10.03.2000].
11. CIMI, The CIMI Z39.50 Profile. Available at URL < http://www.cimi.org/standards/index.html#THREE> [referenced 10.03.2000>
12. Ryan, N. and Smith, D. Database Systems Engineering, International Thompson Computer Press, June 1995.
13. Salton, S. Automatic text processing, the transformation, analysis, and retrieval of information by computer, Reading, Mass., Addison-Wesley, 1988
14. Sheth, A. and Klas, W. (eds.), Multimedia Data Management – using metadata to integrate and apply digital media, McGraw Hill, 1998.
15. W3C, Resource Description Framework (RDF), available at URL < http://www.w3.org/RDF/> [referenced 10.03.2000]
16. W3C, Extensible Markup Language (XML), available at URL <http://www.w3.org/XML/> [referenced 10.03.2000]
17. Weibel. S. et al., RFC 2413: Dublin Core Metadata for Resource Discovery, September 1998. Available at URL <http://www.ietf.org/rfc/> [referenced 10.03.2000]