

RCLIS: towards a digital library for Information Science

José Manuel Barrueco Cruz

University of Valencia (Spain)

Barrueco@uv.es

Imma Subirats Coll

Government of Catalonia (Spain)

Immasubirats@myrealbox.com

1.- Introduction

In this paper we present a case study of a digital library specialised in Information Science: RCLIS. Our aim is to describe the main characteristics of a project in which we have been working for more than three years. RCLIS (Research in Computing, Library and Information Science) is an international co-operative effort to develop a digital library for Information Science. The initiative has two main objectives. Firstly, it tries to compile and to place in the public domain metadata about research documents. The data is freely available for public and private, commercial and no-commercial, purposes. It will also serve as a testbed for digital library research. Secondly to facilitate the access to the freely documents available on the Internet, in order to increase their visibility. RCLIS deals with traditional documents like conference proceedings, articles published in journals and research reports.

RCLIS is inspired by the open source software movement. We believe that closed databases, that are tightly controlled by its vendors, are the equivalent in the world of data to what closed-source software is in the world of software. Users of such databases have to take their contents as given. No formal ways are defined to correct contents or to add records. In the world of academic data, where the user and contributor community closely overlap, it appears that one can improve over closed databases. RCLIS tries to test this idea.

RCLIS offers support to the movement for free online scholarship (FOS) <<http://www.earlham.edu/%7Epeters/fos/>>. It will primarily seeks to document resources that are freely available online. The simple practical reason is that freely available online resources offer more convenience to the user. It will, however, not reject off-line or toll-gated resources because RCLIS aspires to win the trust of all stakeholders in the scholarly communication process.

RCLIS is based in the collaboration of a team of volunteers from different countries. There is not a formal structure or funding from universities or government bodies. Again, like the free software movement, the

work is carried out by a team of friends that work together just for fun. The team co-ordination is done through a discussion list.

The remain of this paper is organised in five sections. Section 2 explains the RCLIS architecture. Section 3 describes weak points detected in the RCLIS model and how they could be fixed. One of these solutions is described in section 4: to implement a new open archive to expand the scope of the digital library into the self-archiving and OAi movements. This archive is called E-LIS. Finally, section 5 concludes the paper.

2.- RCLIS architecture

RCLIS follows the model defined by RePEc (Research Papers in Economics) <<http://repec.org>> . This highly successful digital library specialised in Economics was established in 1997. At the moment it holds metadata about more than 177.000 documents (working papers and articles published in journals). 86.000 of them have the full text available electronically. Data is contributed by more than 250 institutions worldwide and can be accessed using one of the 14 user services available. Last november RePEc got more than 1 million abstract views and more than 212.000 documents were downloaded. More information about RePEc activity is available at: <<http://logec.hhs.se>>. Basically RCLIS tries to import into our discipline the model that has been probed successful in Economics.

RePEc and RCLIS are built on a distributed architecture. They are based in the principle of cost sharing between as many participants as possible, so that each one contribute only a tiny fraction of the work needed to carry out the objectives. Participants in the digital library may be classified in two categories: data and service providers. While data providers (archives) hold metadata about documents, service providers take such metadata in order to provide some added value and to make the result useful for the final user. The interchange of metadata from archives to services is done using a basic set of rules that are specified in a document called *Guildford Protocol (GP)* (Krichel, 1997). Metadata is encoded using a bibliographic format called ReDIF (Research Documents Information Format) (Krichel, 1997). In this way a graphical description of the RCLIS architecture is shown in *figure 1*.

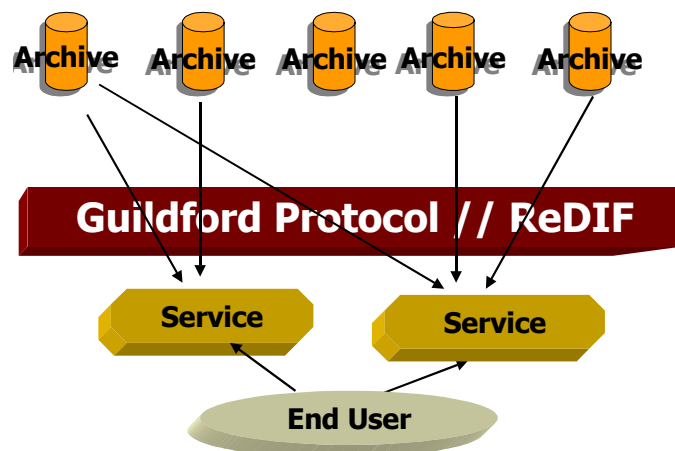


Figure 1.- RCLIS architecture

ReDIF uses a simple structure of field name and field content, which is common and well known by the most part of users. A ReDIF record for this paper will look like:

Template-Type : ReDIF-Paper 1.0
Title: RCLIS: towards a digital library for Information Science
Author-Name: Subirats Coll, Imma
Author-Email: immasubirats@myrealbox.com
Author-Name: Barrueco Cruz, José Manuel
Author-Email: barrueco@uv.es
Keywords: DIGITAL LIBRAIRES; ELECTRONIC PUBLISHING; OPEN ARCHIVES; EPRINTS
Length: 14 pages
Abstract: In this presentation we describe the RCLIS digital library for Library, Information Science and Computing
Creation-Date: 09-01-2003
File-URL: <http://www.uv.es/=barrueco/rclis.pdf>
File-Format: text/html
Handle: RCLIS:nbr:nberwo:6490

The record starts with the specification of the object being described. ReDIF has a broad scope and allows to represent not just documents but all objects involved in the scholarly communication process like authors, research institutions or publication channels (journals, conference proceedings, report series). Then, it follows a set of optional fields including the url of the document full text, if available. The last field is the document handle, which identify each item. It is made up of the string "RCLIS:", a unique code for the archive which provides the metadata, a six letters code that identifies the publication channel and an identification of the item being described. A colon separates each part. RCLIS does not use jet all ReDIF capabilities because is focused only in documents. For an example of full utilisation of the ReDIF format consult Krichel, 2000.

2.1.- Archives and Services

Using RCLIS terminology, data providers are **archives**. Archives are institutions that contribute metadata about the documents they publish or distribute. They provide authoritative metadata in the sense that there is a unique bibliographic description for each document. The creator elaborates such description when the document is made available electronically. That is, it is created at the publication level. Unfortunately, it is not always possible to obtain the collaboration of content providers. Then, it is necessary to relay in third part providers. They describe critical documents for the discipline coming from institutions that are not

participating in RCLIS. Such archives don't provide authoritative data and will be cause of multiple problems (for instance, duplicate records).

From a technical point of view an archive is just a structure of directories and files defined in an FTP or HTTP server. The archive holds static ASCII files containing metadata about documents in ReDIF format. At the moment RCLIS has three main data providers or archives. A handle made up of the string "RCLIS:" and a three letters code identifies each archive:

- **RCLIS:jul** maintained by Julio Alonso Arévalo, librarian at the University of Salamanca (Spain). This archive contains metadata of more than 8943 items (both articles and conference papers). It provides metadata about the most important journals in the discipline like JASIS or Journal of Documentation. Most of them don't have a link to the full text since they are not available in the public domain. In this case, the inclusion of links to the articles full text using OpenURL (Van de Sompel, 2001) is being considered.
- **RCLIS:upv** Maintained at the Polytechnic University of Valencia (Spain). Contains information about more than 600 papers. All documents are available freely on the Internet. It provides metadata about the principal electronic journals of the discipline like Journal of Electronic Publishing, Ariadne or D-Lib.
- **RCLIS:aib** Is an archive hosted by the AIB (Associazione Italiana Biblioteche) and managed by Antonella de Robbio. It holds metadata about journals and conferences organised by the society.

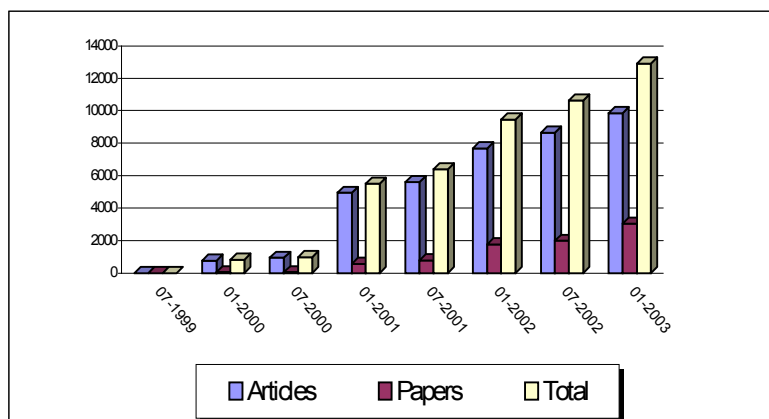


Figure 2.- Evolution of the number of documents in RCLIS

At the moment of writing, December 27, RCLIS holds metadata about 12912 documents, 6852 of them in electronic format and freely available on the Internet. The evolution of the number of documents in RCLIS is shown in figure 2.

Data in ReDIF format, as it is stored in archives, is of little interest to the research community. Initiatives to take the data out of the archives and to give it some type of added value in order to present the information to the end users are needed. This sort of content aggregators is what we call user services. At the moment the main user service for RCLIS is DoIS (Documents in Information Science) available at the url <<http://dois.mimas.ac.uk>>.

DoIS, which was open in 1999, presents the whole data set using static html pages. In this way, the site is fully visible for web crawlers and robots. Access to the site is done using a shwiss++ search engine or alternatively using a browsing facility that allows the user to select the publication type and then a particular channel (journal, conference proceedings) where the single articles are sorted using a chronological criteria. The number of DoIS users has been increasing since the beginning of the project as it is shown in figure 3. Boxes represent number of hits received each month. Line shows the evolution of the hits per document. Since July 2001 this average fluctuates between 16 and 22.

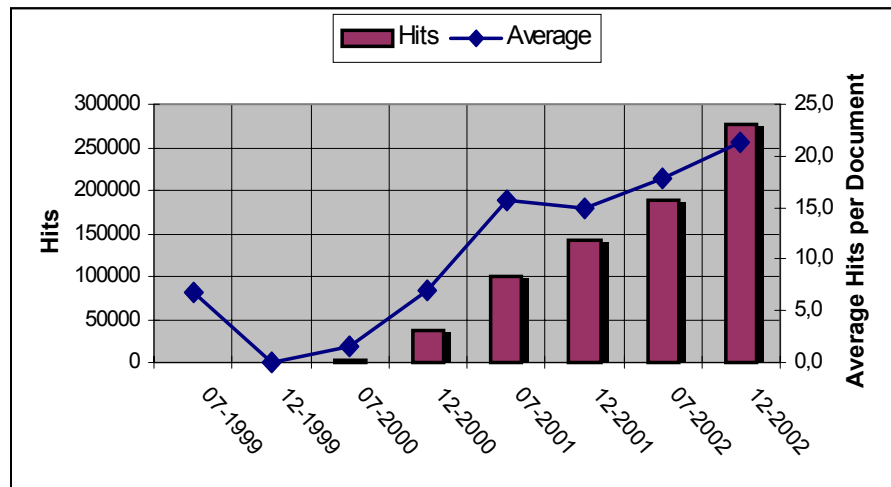


Figure 3.- DoIS usage

3.- Is the RCLIS model working properly? Does it needs any improvement?

A conclusion we could get of three years running is that RCLIS has failed to get on board the main content providers of the discipline. Unlike the master model in Economics, it still relays in a small set of non-authoritative data providers. Why? Because it is not possible to export just the architecture, without taking into account the social issues that surround the research community.

LIS and Economics have very different publication structures. While in Economics there is a tradition of pre-prints and working papers distribution by research institutions, in LIS such tradition doesn't exist. In this way,

in Economics the creation and distribution of contents is much more shared between small players than in other disciplines. Such defragmentation in the distribution of research results forces departments to publicity their pre-print series. In order to obtain visibility in the community. An international digital library like RePEc makes the advertising easier, therefore departments are whiling to contribute to a project like RePEc.

On the other side, publication in LIS is concentrated in journals and conference proceedings. While the former are published by commercial companies, the later are organised by large societies and institutions. None of them are proud to contribute to an effort like RCLIS if they don't see clearly the advantages of doing so. Even in Economics, metadata from the big publishers has been obtained after hard negotiations and is currently maintained by third part institutions.

To sum up, if RCLIS has failed to get on board the main content providers, it is the moment of looking for new data input methods. A first solution came from changing the object with which we are dealing. Instead of dealing with publishers alone, we should deal directly with the authors themselves, asking them to store in RCLIS electronic versions of the documents they publish. That is, to show them how to self-archive their publications. Why should they be interested in self-archiving? Because self-archiving their works they could get more visibility for them. It has been probed that documents made available freely on the Internet are cited more often than those that are hidden under commercial web sites (Lawrence, 2001).

There is an international movement to free the scholarly literature lead by Stevan Harnad (Harnad, 2001). The self-archiving initiative is based in the idea that authors are the only owners of the documents they produce and could store them in their home pages or in public archives. Such documents are *e-prints*, a new term that designs both documents that have passed a quality control or peer review process (post-prints) and those that haven't passed such quality control (pre-prints). In this way the scientific community benefit of the free access to the scholarly literature and authors get a higher visibility for their works since they aren't economical barriers to such access.

But documents in home pages or isolated web servers are of little interest if they can't be discovered by third part content aggregators. The OA initiative <<http://www.openarchives.org>>, came to solve this problem. OAi was created to allow the federation of content providers so that they could interoperate and interchange metadata on the Internet. It differentiates between data and service providers like RCLIS does (Note that a member of the RCLIS team, Thomas Krichel, has been in the technical committee of the OAi since the Santa Fe Convention). There are two types of archives: discipline and institution based. The first one holds metadata about documents in a concrete subject area but coming from multiple institutions. The second one is an institution, which holds metadata about multiple disciplines but with the common denominator of being published by its staff. In the RCLIS case, it was decided to create a discipline-based archive. In this way in January 2003 was created E-LIS (Eprints in Library and Information Science). <<http://eprints.rclis.org>>.

4.- E-LIS an open archive for our discipline

E-LIS has been designed as an international open access archive for eprints on Library, Information Science and related disciplines. Its purpose is to make the full text of scientific documents visible, accessible, harvestable, searchable and useable by any potential user with access to the Internet. Furthermore this service aims to support individuals who wish to publish or otherwise make their papers (published or not) available world-wide.

Searching and archiving in E-LIS are totally free for any user. The only requirement is that authors wishing to submit a document need to register in order to obtain a user id in the system. Librarians, libraries, research institutes, organisations and individual researchers involved in LIS and related fields are encouraged to make use and contribute to the archive.

4.1 Similar initiatives

E-LIS is not neither the only or the newer archive. There are two other initiatives to create open archives for our discipline:

- **@rchiveSIC** <<http://archivesic.ccsd.cnrs.fr/>> is a French collaboration project between several research institutions. At the moment it holds about 80 documents, the most part of them in French. Access to the papers is done through a subject classification scheme of 22 categories. From this classification we can see that the scope of the archive is not just libraries but related disciplines too (Museology, Education, Ecology, etc). There is no information in the web site about their submission policy or copyright restrictions.
- **DLIST** (Digital Library of Information Science and Technology) <<http://dlist.sir.arizona.edu/>>. It is a service of the School of Information Resources and Library Science and Arizona Health Sciences Library (University of Arizona). At the moment the archive stores more than 100 documents. The access is through a detailed list of topics. The archive aim is to store all types of scholarly documents in Information Science but with two subject areas of emphasis: information literacy (educational materials like tutorials, etc.) and informetrics. The deposit could be done by the author herself or by the archive if the author submit by email the document. They accept only documents in English.

Only @rchiveSIC is currently registered as OAI data provider.

The question could be, why another archive for LIS is needed. E-LIS does not try to compete with the established initiatives but provide alternative possibilities to the authors. Additionally there is a geographical question too. In this sense @rchiveSIC is centred in France while DLIST is centred in USA and English

documents. There is a need for a true international effort that deal with all documents without language or geographic restrictions.

4.2 The eprints software

As other initiatives listed in the previous section, E-LIS has been built using the eprints software <<http://www.eprints.org>>. Eprints has been developed at the University of Southampton. It is a popular system to implement open archives, which it is being used by more than 30 repositories. Eprints has been designed with the main objective of being easy and fast to install and, of course, freely distributed. In fact eprints is made available under the GNU license, which means that the source code is available in the public domain and could be used for anybody. The main characteristics of this software are:

- Simplicity of installation and configuration. Nevertheless a system administrator with experience in UNIX and perl programming skills is still needed in order to do the first set up of the software.
- It allows both to store documents in any format, and to deposit a document in several formats. The document submission is done using a very simple web interface.
- There is flexibility in the metadata format used to describe the documents. The system provides a simple element set that could be expanded if the institution needs a more detailed format.

4.3 Contribution policy

In order to get the maximum number of authors on board, the contribution policy is very simple. In a broad sense any document related by topic with LIS and available electronically in any format could be included in the archive. The basic criteria for acceptance is that documents must be relevant to research in LIS fields and they should have the form of a finished document that is ready to enter into a process of scholarly communication. That doesn't mean there is no an edition procedure to make sure that authors do not submit garbage or inappropriate content. An editorial board made up of researchers in different areas of our discipline is in charged of examining the documents submitted. The workflow for a typical submission could be as follows:

1. New author registers via a web interface. When registering the author enters metadata about herself like contact addresses or subjects of interest. He can also subscribe to a mailing list where new additions to the archive are announced.
2. Author submits an eprint. The process takes two basic steps:
 - Creation of metadata about the document via web forms. The number of fields to fill in depends on the type of document being submitted. Only a few fields are mandatory. The archive admits all

document types. If a paper does not match any of the categories provided, the author can ask the editor to include new types.

- Uploading of the document full text. The file containing the full text can be located either in the local author's machine or in any server accessible on the Internet. In this last case a url to the document is required.
3. Submitted documents are placed into a buffer where they are examined by a member of the editorial board who can approve the submission, reject it outright or return it to the author for modifications.
 4. When the editorial board has approved the eprint it is included in the archive and can be accessed via the search engine or via the browsing facilities (year or subject). Each eprint is described in an html page that includes a url to the full text. The eprint is now ready to be harvestable via the OAI interface too.

There are not restrictions in the file format used, nevertheless PDF documents are strongly recommended. The archive accepts postings in all languages too, but an abstract in English is required when the text is in a different language. The copyright issues about the documents being submitted are very important. In this sense, submitting authors are responsible of being sure the documents they archive haven't any copyright restriction in their electronic distribution. They are asked to not submit publisher produced PDF or other format versions. Unless noted otherwise, the creators or authors retain copyright and other proprietary rights.

Finally, It is mandatory that the depositor should be the author or one of the authors of the deposited work. The editor will verify this and reject the submission otherwise.

4.4 Structure

The archive structure is made up of three main parts: the access module, the internal database and the user area. Like other services built on the Eprints software, E-LIS is accessible in two complementary ways. Firstly a search engine is provided in order to seek the bibliographic descriptions. Secondly the user can browse eprints by creation year and subject. When submitting a document is mandatory to include the date of publication and to assign it one or several classification codes.

The subject tree adopted is named JITA Classification Scheme <<http://rclis.org/internal/jita.txt>>. It has been built for E-LIS on the basis of NewsAgent Topic Classification Scheme <<http://users.aber.ac.uk/emk/topics.htm>> and RIS Classification Scheme <<http://www.iud.fh-darmstadt.de/iud/wwwmeth/publ/example/werkz/risclass/menu1.htm>>. JITA's objective is to provide a simple subject schema to categorise the most part of documents in the discipline. The scheme is open. In order to

keep simplicity there is only a single level of categories, but the scheme is ready to incorporate more specific levels to the tree when needed.

- A. Theoretical and general aspects of libraries and information
- B. Information use and sociology of information
- C. Users, literacy and reading
- D. Libraries as physical collections
- E. Publishing and legal issues
- F. Management
- G. Industry, profession and education
- H. Information sources, supports, and channels
- I. Information treatment for information services
- J. Technical services in libraries, archives and museums
- K. Housing technologies
- L. Information technology and library technology

The internal database works on a MySQL database management system. All metadata about eprints and users, and all information required for the archive administration is stored in SQL tables.

The user area is made up of a set of perl scripts that read directly the information stored in the MySQL database. For this reason the html pages are created dynamically, on the fly. The first step to enter the personal area, once the author has registered, is to give an username and password. The author is presented then with a menu where the main option is to add an eprint to the archive or modify the documents already stored. Additional features available at the user area are: modify or complete the author metadata, subscribe for getting email alerts about new documents, list the author's documents already in the archive, searches for users, etc.

6.- Conclusion

In this paper we have described the RCLIS digital library for Information Science. There have been two parts in the live of the project. Until this year we have been working with an architectural model borrowed from the Economics discipline where it has been up for almost five years now. Since in Information Science the publication structure of research results is quite different than in Economics, new ways to obtain input of documents need to be investigated. This year, RCLIS has moved into the OAi movement by creating an open archive for our discipline. In this way, authors could self-archive their works when not available in the public domain. With this new development the new architecture of RCLIS is shown in figure 4.

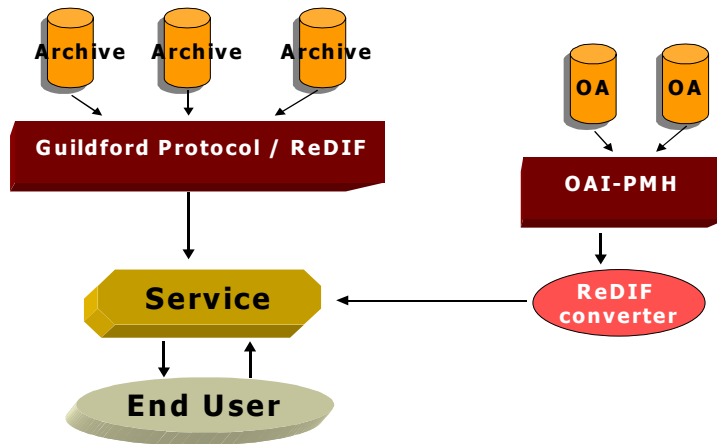


Figure 4.- New RCLIS architecture

In this graph we can see how a new data input has been included to allow services to harvest both metadata from RCLIS archives which use a combination of GP and ReDIF and archives like E-LIS that uses the Open Archives Protocol for Metadata Harvesting (Lagoze, 2002). In this last case a converter from the original metadata in the archives, that is distributed using the unqualified Dublin Core format, is needed.

To sum up, RCLIS is in a transition phase where two types of archives and protocols live together. We hope in the near future the old archives based in GP will be moving forward to adopt the OAI-PMH model.

References

- Harnad, Stevan. 2001. The self-archiving initiative: Freeing the refereed research literature online. *Nature* 410, 26 April 2001, pp 1024 – 1025
- Krichel, Thomas (ed.) 1997. Guildford Protocol. Available at: <ftp://netec.mcc.ac.uk/pub/NetEc/RePEc/all/root/docu/guilp.html>
- Krichel, Thomas (ed.) 1997. Research Documents Information Format. Available at: ftp://netec.mcc.ac.uk/pub/NetEc/RePEc/all/root/docu/redif_1.html
- Krichel, Thomas. 2000. RePEc, an open library for Economics. Forthcoming in a book by MIT Press. Available at: <http://openlib.org/home/krichel/papers/salisbury.html>
- Lagoze, Carl et al. (eds.) 2002. The Open Archives Initiative Protocol for Metadata Harvesting. Available at: <http://www.openarchives.org/OAI/openarchivesprotocol.html>.
- Van de Sompel, Herbert and Beit-Arie, Oren. 2001. Open Linking in the Scholarly Information Environment Using the OpenURL Framework. *D-Lib*, vol 7, no. 3. Available at: <http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html>.
- Lawrence, Steve. 2001. Online or invisible? *Nature*, vol. 411, no. 6837, p. 521. Available at: <http://www.neci.nec.com/~lawrence/papers/online-nature01/>.