Experience in Creating an Electronic Publications Archive at a National Scientific Laboratory

Matthew Mascord, Catherine Jones, Simon Lambert
CCLRC

Summary

This paper describes the development of an electronic publications archive for CCLRC, a national scientific laboratory in the UK. The need for such an organisation-wide archive is pressing, yet for historical reasons individual departments and facilities have their own customs and practices, and these have to be respected in order to ensure acceptance of the system. It thus serves as a real-world case study of the issues in developing such an archive—issues that are both technical and concerned with policies. The paper outlines these issues and the solutions that are being adopted.

1 Introduction

The Council for the Central Laboratory of the Research Councils is a Government funded Non-Departmental Public Body and is a Research Council in its own right. It has three remits: providing large-scale scientific facilities for the UK and international community, widening knowledge by performing research, and providing access to this information for the general public. CCLRC is based on three sites: Chilbolton in Hampshire; Daresbury Laboratory in Cheshire and Rutherford Appleton Laboratory in Oxfordshire.

CCLRC is made up of both departments and specialist centres in a federal manner. The major departments provide facilities such as the Neutron Spallation Source (ISIS), a range of Lasers and a Synchrotron. Other departments do their research using facilities such as CERN, the European Synchrotron Radiation Facility (ESRF) or space-based facilities such as the Mars Explorer. A large amount of the research performed at CCLRC is based on collaborations with universities and other organisations, especially in the Particle Physics area, and this means that scientific papers can have hundreds of authors in some cases.

The main form of dissemination is by publishing in a peer-reviewed journal or conference proceedings. However CCLRC does publish using its own CCLRC Reports sequences which are approved internally. There are four types: Conference proceedings, Preprints, Technical reports and Theses. The preprint series were

used extensively in the past for publicising new scientific developments and have been mostly superseded by using electronic preprint archives.

The fact that the organisation has a federal structure means that there are wide differences in operating practices between the departments., for example in the preferred means of publication of results. CCLRC differs from UK universities in that it has no teaching function, and is not subject to the Research Assessment Exercise. However it is judged by the quality of its scientific output, and publications are obviously the key to this.

2 The need for an electronic publications archive

To be able to assess the scientific output of the organisation it is necessary to have the information in one place. As CCLRC is organised federally there are collections based within departments but until now no overall collection, which has complicated the process of assembling publication lists and resulted in needless fragmentation. Therefore a project has been set up to develop an electronic publications archive for CCLRC.

The feasibility study established the need for the publication archive to widen the remit from CCLRC staff publications to publications that had used CCLRC facilities when doing the research. As CCLRC provides these large-scale experiments then limiting the scientific output to only CCLRC staff excludes the greater part of the research done here.

It is very difficult, without this project, to ascertain the total number of publications. The feasibility study looked at identifying CCLRC publications using external sources such as Web of Science and preprint archives such as CERN CDS & ArXiv. A conservative estimate is 1000 publications per year.

Once a central publications archive is in place then it can be used for a variety of purposes, some of the most important being:
Producing departmental publication lists: Most departments produce an annual report with scientific highlights and a publications list. This list is an important part of the reporting cycle and is very tedious to achieve.
Individual publication lists: People who publish papers need to keep an up to date list and this system will provide a facility for doing this.
Performance indicators: CCLRC and individual departments need to show that they are providing value for the public money spent here. A good indicator of scientific worth is the number of publications produced. This system will be able to provide this information at many different levels.

The feasibility study identified a need for a centralised publication archive but also identified a number of areas where there might be problems. The differences in working practices between the departments meant that the design needed to be as non-prescriptive as possible. Another major hurdle was how the data was to be entered and what incentives there were for authors to submit works themselves. We concluded that there will be some authors who will be happy to enter details themselves as it will result in their up-to-date publication list, there will be others that co-operate as part of a procedure to produce departmental lists and there will be some authors who will expect some central co-ordinator to do it for them.

Although the original remit of the project was not to include historical data one department's existing database was loaded into the prototype and this encouraged the project sponsors to see the benefits. It was therefore decided to actively seek historical information from all publishing departments. Information has been received from every department approached and whilst this has made the project longer, it has been worth the extension as the potential users can see the benefits of a centralised system.

3 Technical options and considerations

Archive or organisational bibliography?
For work that CCLRC publishes *internally*, under its own name, the ePublication Archive must store the definitive fulltext of the material. However for work that CCLRC staff and facility users publish *externally* its function must be to hold publication information (navigational metadata) to link to the definitive fulltext stored externally whether this is at an external ePrint archive such as arXiv.org [Cornell University, 2003], e-Journal such as BioMedCentral [BioMed Central, 2003] or traditional printed journal issue or conference proceedings. In this way the CCLRC ePublication Archive must act simultaneously as both *archive* and *organisational bibliography*.

Relevant standards
There are a wide variety of metadata standards that could be adopted by a digital library. MARC [Library of Congress, 2003] is very standard but complex, targeted at print and limited in its ability to describe serials at the article level. Dublin Core Library Application Profile [Dublin Core Metadata Initiative, 2002] provides a standard set of fields to describe resources within a digital library but it is not machine understandable and is limited in its ability to describe *where* a work has been published. IFLA FRBR [IFLA, 1998] provides a clear conceptual framework grouping records under overarching Works, Expressions, Manifestations and Items. ONIX for Serials [Editeur, 2002], aimed at serials publishers, gives a detailed XML

schema for describing individual Serial Volumes, Issues & Articles. BiBTeX [Patashnik, 1988] is a de facto standard used by the LaTeX document preparation system for storing bibliographic references.

Off the shelf packages
There are a number of off-the-shelf digital library software packages. The two leading Open Source systems were considered: MIT-HP DSpace [MIT-HP, 2003] and GNU Eprints [University of Southampton, 2003]. GNU Eprints produced by the University of Southampton provides a low barrier route for institutions to create online archives and aims to maximise research impact through self-archiving. DSpace is a digital library system developed jointly by MIT and Hewlett Packard and provides a stable platform for capturing and preserving an institution's digital resources.

Both support the metadata harvesting protocol from the Open Archives Initiative (OAI-PMH) that promises an application independent interoperability framework [Open Archives Initiative, 2003] allowing multiple distributed archives to operate as one entity.

Dspace and GNU Eprints both use DC Library Application Profile as their native metadata format. The unstructured nature of DC makes it difficult to unambiguously link a Work to CERIF entities such as Project, Person and Organisational Unit crucial for the production of individual and organisational bibliographies, performance indicators and links to external systems such as records management and the CCLRC data portal.

A bespoke system
The alternative to using an off-the-shelf system is to develop a bespoke system. The aim would then be to develop the system using Open Standards, and in order to minimize the amount of programming to construct the system from well defined black box components. The following options were examined: JSP/JSTL, Cocoon (XML/XSLT), Struts, Velocity & Turbine as potential web frameworks, Apache Lucene as search engine, OAICat for OAI interface and PostgreSQL, Oracle and MS SQL as candidate RDBMSs.

Historical data
The desire to link metadata to CERIF entities is generally more important for current and future research activities than it is for research activities that occurred in the past. Therefore the benefits of restructuring retrospective data to comply with a very structured model may not be justified by the additional cost. Nevertheless the feasibility study found that having the data available for search and browse albeit in a descriptive freetext form is of great benefit. This introduces a need to cater for different levels of metadata structure within the system.

Functionality

The feasibility study identified a number of functionalities that should be supported by a publications archive. For clarity they were divided into three areas: publication tracking, publication approval and archive.

The publication tracking functionalities were the functionalities that departments require to keep track of what and where they are publishing, the publication approval for departments to ensure that what they publish is of an acceptable standard and finally the archive to allow interested enquirers to search and browse.

Publication tracking

In the publication tracking category there are two classes of users: authors and departmental administrators. For departments to report on what exactly they publish, authors must notify them. The system should facilitate this by providing some form of lightweight web form that authors can complete. To encourage participation, the form should have an extremely shallow learning curve and not force the user to enter overly structured data, e.g. by providing freetext fields as fallback options. The important thing being that the author *identifies* the publication and not that the metadata is highly structured.

Departmental administrators are interested in compiling publication lists. To do this they need to be able to unambiguously identify papers written by members of their department or users in the case of facility department: this implies some form of name authority control. Another complication is papers that are written as collaborations between two or more departments. Given the diverse and potentially incomplete nature of the metadata entered by authors the departmental administrators would also need facilities to carry out a degree of metadata validation and enrichment. Finally there would need to be the ability to detect and remove duplicates that may have been added by two or more authors of the same paper.

Publication approval

In the publication approval category there is a need for some sort of approval workflow that allows publications uploaded to the system to remain in an intermediate state invisible to external enquirers before being approved and subsequently visible to the outside world. It is possible to envisage the need for automated emails to be sent to approvers when something is submitted and for a workspace to allow an approver to view all submissions pending approval. This is the strategy implemented by the Dspace system.

Publication archive

In the archive category there is a need for a flexible resource discovery framework: an ability to browse the archive by various categories such as Year, Author,

Department, Facility a quick search and then an advanced search for more experienced users. In this way the user has multiple routes to a particular publication.

4 Development of the publications archive

Draft data model
On completion of the feasibility study the first activity to be undertaken was the construction of a draft data model and mocking up of HTML pages to look like a potential real system yet not connected to any underlying database. These helped to analyse the data requirements whilst considering usability issues.

A formalized version of Dublin Core, proposed by Keith G Jeffery, was taken as the starting point but then significantly extended with elements taken from other metadata standards. One of the ways in which DC was found limiting was that a separate bibliographic record has to be created for each and every manifestation of a Work. We wanted to be able identify the creative Work independently of Manifestation and also to attach metadata at the Work level as it generally does not change across manifestations. To do this we adopted the Work and Manifestation elements of the IFLA FRBR model.

As DC sets out to describe a specific instance of a digital resource as opposed to its publication context it does not provide well for the identification of journal articles, conference events, book chapters etc. Therefore we adopted a subset ONIX for Serials for this purpose and adapted it to a relational model. Finally we took elements from BibTeX where FRBR, DC and ONIX were insufficient. The draft data model was then iteratively refined taking into account real data stored in systems such as arXiv.org and existing CCLRC publication databases.

Initial prototype
As an initial prototype of the draft data model and screenshots PostgreSQL, JSP and JSTL were chosen. Shortly afterwards we decided to migrate to PostgreSQL, XSLT/XML and Cocoon due to limitations with JSTL—for example the SQL library with JSTL is useful for prototyping but is not scalable. Cocoon also provides a declarative mechanism for moving data from an RDBMS through XML and finally to a variety of presentation formats including HTML. To make the system look more like a real system we loaded in the 8000 publications dating back to 1988 from the publications database of ISIS, one of the major facilities at CCLRC.

Demonstrations
In order to begin showing the system to users everyone from the feasibility study was contacted, offering to show them the prototype. Having the real data within the

system generated interest and other departments quickly offered their data. In a short space of time there were around 15000 publication records from five departments and facilities of CCLRC. Writing migration scripts for such large quantities of data proved to be more time consuming than anticipated and this slowed the development of the archive. Not unexpectedly a number of modifications to the data model were also required.

Migration to Oracle

The existing database implementation of PostgreSQL running on Cygwin on Win32 was proving to be inadequate to deal with the additional volumes of data that were being imported into the system. This was mainly because PostgreSQL by default is configured to use the minimum of system resources and it is difficult on the Cygwin to change this. As CCLRC already owned a site wide Oracle licence we chose to migrate to Oracle 9iR2.

Current status

The current status of the system is read only and linked to CCLRC's internal library web pages. It is not currently accessible on the WWW. There are around 20000 bibliographic records covering eight of CCLRC's departments. ISIS have established an ODBC link with their Access database to allow the new system to hold all future bibliographic records.

The system provides various browse indices, quick search and also an advanced search. Many of the particle physics papers also have multiple links to freetext versions made possible by the data from SPIRES.

Future plans

In future it is planned to continue improving the quality of the metadata and to move the remaining CCLRC departments onto the publications archive. There is a lot of potential for using the CrossRef system to link to electronic copies of journal articles using DOI. Another interesting possibility is the linking of the publications archive to the eScience CCLRC data portal. This would allow publications to be linked to actual data from scientific experiments and vice versa.

Finally we hope that the system will eventually be used by scientists as a resource discovery tool in its own right. To make this a reality we are looking at ways of improving resource discovery by using automated thesauri. We plan on providing RDF feeds that can be linked to an RDF thesaurus such as the one being developed by the SWAD Europe project [SWAD-Europe Thesaurus Activity, 2004].

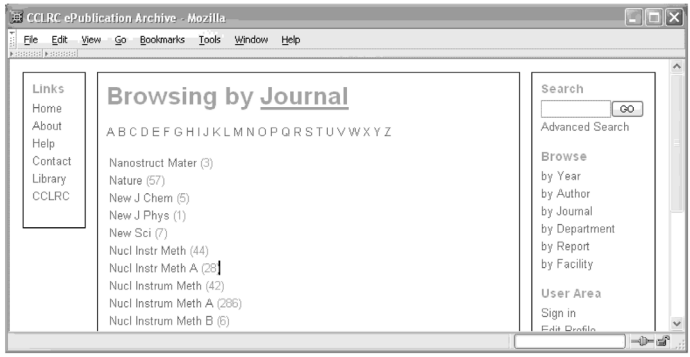

Figure 4.1 – Browsing by journal

Figure 4.2 – Browsing by author

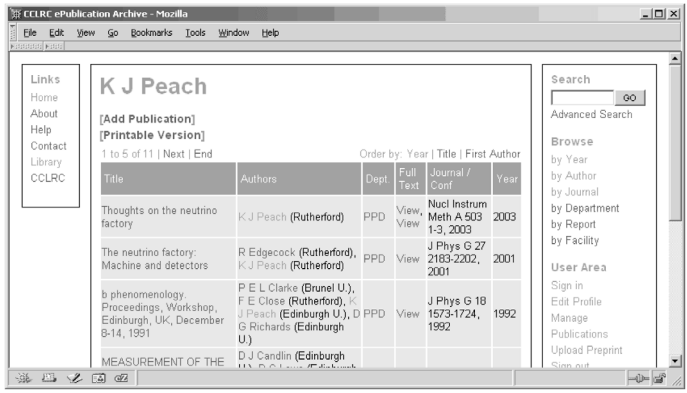

Figure 4.3 – Publication list for an individual

5 Policy issues

Our project sponsors envisaged that each author would use the publications archive to add their own information. The feasibility study showed that in facility departments there is a lot of effort expended to produce comprehensive publications lists that include non-CCLRC publications. The custom of recognising CCLRC scientists who helped with an experiment but did not analyse the results leads to them being authors on publications they know nothing about. These complications mean that we have identified three classes of people who are likely to input into the system.

Authors: There will be some people, especially those who use bibliographic software at present who will input their information into the system

Department co-ordinators: These people already exist in areas which produce annual reports. They will encourage people to self-submit, they will add external publications in and ensure that the publications for their department is as complete as possible

Administrative support: Once the system is available externally then there needs to be some quality assurance such as spelling checking. It is envisaged that the Library and Information Service will fulfil this role.

One of the major policy issues in this project has been one of the original aims that sought to provide a process to ensure that publications affiliated to CCLRC had been quality checked. Although CCLRC does have a formal publication approval policy, this is devolved to the departments and it has been interpreted in different ways. One of our project sponsors operates a very formal approval system but most of the other departments use line managers and the peer review system to ensure appropriate quality. To some potential users a firmer approach to approval is an encroachment on scientific freedom and is to be resisted. As the project progressed it was decided that the approval aspect would be dealt with outside the remit of the system, leaving existing approvals procedures in place as it would not be possible to achieve an harmonised approach.

At present the system is being developed with the ability to upload the preprint and to link to published versions of the article on publishers websites. This is due to copyright restrictions. The copyright of the former is owned by CCLRC and access to the latter is due to journal subscriptions taken out by CCLRC Library and Information Service. This area is one where there are large changes to the practice happening at present. Many publishers allow the author to put the full text of a published article on the institutions web site for internal use. We feel that this is an area which will change greatly over the next few years.

6 Conclusion

A number of important lessons have been learnt in the development of the electronic publications archive. No off-the-shelf solution is quite suitable for the purposes required, and thus a bespoke system is being developed, using existing components where possible. A really key issue is whether and how to incorporate historical publications data: this undoubtedly enhances the utility of the system, but the effort involved in incorporating this data should not be under-estimated. Furthermore, the historical data is of different levels of quality and completeness, and decisions are needed on how to handle this aspect.

It is also recognised that to link publications data with CERIF entities, the metadata has to be much more structured than Dublin Core permits.

7 References

Cornell University (2003): arXiv.org
BioMed Central (2003): BioMed Central. www.biomedcentral.com
Dublin Core Metadata Initiative (2002): Library Application Profile. dublincore.org/documents/library-application-profile/
Library of Congress (2003): MARC. www.loc.gov/marc/
IFLA Functional Requirements for Bibliographic Records (1998): www.ifla.org/
Editeur (2002): ONIX for Serials. www.editeur.org/onixserials.html
Patashnik, Oren (1988): BibTeXing. www.ecst.csuchico.edu/~jacobsd/bib/formats/bibtex.html
MIT-HP (2003): DSpace. www.dspace.org
University of Southampton (2003): GNU Eprints. www.eprints.org
Open Archives Initiative (2003): Protocol for Metadata Harvesting. www.openarchives.org/OAI/openarchivesprotocol.html
SWAD-Europe Thesaurus Activity (2004): RDF Thesaurus. www.w3c.rl.ac.uk/SWAD/rdfthes.html

8 Contact Information

Matthew Mascord
Business and Information Technology Department
CCLRC Rutherford Appleton Laboratory
Chilton, Didcot
Oxfordshire OX11 0QX
UK

e-mail: M.Mascord@rl.ac.uk