Digitisation and Preservation of CCLRC's scientific legacy: the ePubs Technical Reports Project

Catherine Jones

CCLRC, Rutherford Appleton Laboratory, Chilton, Didcot, OXON, OX11 0QX, UK
c.m.jones@rl.ac.uk

Abstract. This paper introduces the Council for the Central Laboratory of the Research Councils (CCLRC) and describes the publishing role of a scientific research organization. I discuss the Library and Information Service's archiving role and the strategy and policy for future developments. I introduce the Institutional Repository, ePubs, and explain its use in the digital preservation of Technical Reports. I outline the lessons learnt and the challenges still to be overcome.

1 Introduction

The Council for the Central Laboratory of the Research Councils (CCLRC) is a UK Government funded Non-Departmental Public Body and is one of the eight UK Research Councils. It has 1700 employees, a turnover of £240 million and 360,000 users p.a. CCLRC has three aims: to provide large-scale scientific facilities for the UK and international community, to extend knowledge by performing research, and to engage with the general public. It was formed in 1995 from the existing Chilbolton Observatory in Hampshire; Daresbury Laboratory in Cheshire and Rutherford Appleton Laboratory in Oxfordshire. The constituent parts date back to the 1950s (RAL) and 1960s (DL).

To support CCLRC's aims, there is a need both to keep a record and provide a means of publicising and disseminating technical innovations; which has been fulfilled by Laboratory reports and associated archiving and dissemination managed by the CCLRC Library and Information Service (LIS).

2 CCLRC's Publishing Role

CCLRC and its forebears have always published Laboratory reports. These are formal publications and are placed with the Deposit Libraries. As the publisher CCLRC has rights and responsibilities to curate and preserve this material. At present this is done by traditional archival and librarianship skills using the original paper copies. They fall into four main types:

- **Technical:** these are for important scientific and technical issues that are not suitable for journal articles. This is unique information and is a high priority for

digitization and preservation and has long-term relevance. As the publisher CCLRC owns the copyright to this material.

- **Preprints:** We have a long history of working in Particle Physics and paper preprints were the forerunner to arXiv [1].
- **Theses:** Historically all UK Particle Physics theses were produced as part of the RAL Theses sequence
- **Conference Proceedings:** DL and RAL organized conferences.

The number of CCLRC reports is decreasing due to the changes in publishing and dissemination mechanisms.

It is intended that a more flexible and high profile approach to the electronic publication of lab reports will reverse this trend.

3 CCLRC Library and Information Service Strategy

The CCLRC Library and Information Service (LIS) has a strategy for service development over the next three years which can be summarised by the phrase "*From print warehouse to electronic resource portal*".

One of the areas of "*print warehouse*" is the CCLRC Laboratory report collection which spans 50 years at RAL and 40 at DL taking up 135 metres of shelf space. The collection provides three services: a bound archive for preservation, two loan copies for use by our readers and spare copies which are sent out on request.

A key project for 2005/6 is to digitise the CCLRC Technical Report archive. The benefits will include increased access and visibility of these publications whilst reducing the physical footprint of the collection.

Our Institutional Repository, ePubs, is a key building block in the "*electronic resource portal"*. Its collection remit is the scientific and technical output of CCLRC, the bulk of which is journal articles but includes Laboratory reports. From a Library and Information Services perspective this means that the ePubs system is strategically important to us in the evolving information environment for two reasons. We see Open Access and a network of Institutional Repositories as a means of reducing reliance on journal subscriptions and it is an important asset in the publicity and dissemination of CCLRC produced literature.

3.1 Policy Decisions

After some consideration we have decided to use ePubs as both our dissemination and archival system for Technical Reports; linking the present day use to the archival copy thus reducing the amount of identical bibliographic metadata kept.

– **Policy issues:** Only two copies of any particular report will be kept, one of which will be the bound archive. We will outsource the digitization. We will preserve both the text and the layout, including graphs, formulae and tables. The LIS is

committed to funding ePubs in the long term and will investigate the preservation costs with a view to maintaining funding for this.

- **Technical issues:** TIFF files will be used for the archive copy and PDF for dissemination. The reports will be OCR'd and the text files produced will be used for free text searching within ePubs. We will use the Atlas Petabyte Store for the TIFF files giving it the responsibility for the bit level preservation of the files.

4 Role of the Institutional Repository

CCLRC's Institutional Repository, ePubs [2] has been accessible since May 2004. It contains 21,000 records of which approximately 400 are full text, mostly Laboratory reports. It is an in-house development, using Oracle, Apache Cocoon, Lucene and XSLT/XML. ePubs has an OAI interface conforming to OAI-PMH and is harvested regularly by OAIster, Google and Google Scholar. It has been selected by Thomson ISI for inclusion in the Current Web Contents product in recognition of the high quality data contained within the system.

4.1 Overview

The conceptual basis of the ePubs system is the IFLA Functional Requirements for Bibliographic Records (FRBR) [3] model, with enhancements from ONIX for Serials and Formalised Dublin Core. In FRBR model there are four entities that are used to describe any particular work: the Work is an abstract concept of a distinct intellectual or artist creation. A particular work can be *realised* as an expression, or series of expressions. Each expression will be *physically embodied* as a manifestation, or series of manifestations. A single exemplar of the manifestation is known as an item. ePubs implements the Work, Expression and Manifestation elements at present.

The advantages of this approach mean that we can group the different expressions and associated manifestations within the same concept (metadata record) reducing the number of "identical" records and providing easier retrieval for our users. It is easier to identify an eprint from the published work within a record rather than to have to compare records.

4.2 Adaptations for Preservation

We are merging the concept of an Open Access Repository with an Open Archival Information System (OAIS) [4], in that although we will be using the Submission and Archival Information Package (SIP & AIP) principles, we intend to disseminate the PDF copy of the work to the general users and only give access to the archived copy within an archive manager interface.

We will enhance the submission process to include a new manifestation description. Manifestations are used to describe particular physical formats, such as technical report or journal article. The new description will be an archive copy and

will include preservation description information such as provenance, context, reference and structure, as discussed in Digitising Collections [5].

We are also using the draft of the Trusted Repository Audit Checklist [6] to examine the quality of the repository and its management and to provide feedback to the RLG. Of the four areas covered by the checklist, the organisational and technical levels are well covered, the functional and communication levels are under development. This is a useful tool for ensuring good quality development.

5 Lessons learnt & Challenges to be faced

The main lesson learnt is that this type of project always takes longer than anticipated! In particular, changes to priorities and resourcing make a big impact on research led activities within a small service focused team.

There are two further challenges in this area: a) estimating the cost-effectiveness of attempting digital preservation of a paper resource: will it in fact be cheaper to maintain the archive in paper and reinvest in digitization when technology changes? b) born-digital reports: in this case we start with digital format, such as Word or LaTEX which we know will need managing for the long-term preservation.

6. Conclusions

This project is an interesting one and should enable LIS to investigate the policy and technical issues surrounding the digitization and preservation of technical material whilst maintaining a fall-back position of paper archiving.

References

1. http://arxiv.org/
2. http://epubs.cclrc.ac.uk/
3. IFLA Study Group. : Functional Requirements for bibliographic records: final report. Saur, Munich (1998) http://www.ifla.org/VII/s13/frbr/frbr.pdf
4. http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf
5. Hughes, L.: Digitizing Collections. Facet Publishing, London (2004) 196-198
6. http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf