e-Infrastructure to Support Research Management for Benefit

Keith G Jeffery
Director IT & International Strategy
Science & Technology Facilities Council
keith.g.jeffery@rl.ac.uk

Abstract

The e-infrastructure hides the heterogeneity of the underlying physical ICT (Information and Communications Technology). Upon this base, we can make information about research available. There has been a flourishing of information provision e.g. in html web pages. However, such un- or semi-structured information - while useful for human reading - does not permit precise IT searching and retrieval, nor display and subsequent processing.

The emergence of institutional repositories of scholarly publications - representing the intellectual property of the organisation - has provided open access to much research information. However, the repositories usually utilise a form of metadata (Dublin Core) that does not have a fully formal syntax (and so has variable structure - difficult for automated processing) and does not have defined semantics (meaning).

Work at STFC in UK and more generally within euroCRIS (www.eurocris.org) has demonstrated that research information can be made available optimally by utilising a CERIF (Common European Research Information Format) CRIS (Current Research Information System) as a source of research information and as the formal metadata to describe the content of repositories both of scholarly publications and of research datasets and software. This facility can then be used by researchers to find potential colleagues or to track the research of others, by research institution decision-makers to plan strategic development of research, by funding agencies to assess the output from their funding, by entrepreneurs to find research ideas suitable for exploitation leading to wealth creation / improvement in the quality of life and by the media to inform the public on 'research stories'.

1. Introduction

Research Information is a valuable resource for recording research achievements, for encouraging further research, for stimulating innovation and wealth creation and for the media to use to inform public debate. The real benefit is realised when:

    (a) it has appropriate metadata to improve the accuracy and speed of search, to assist interoperation and to manage security, privacy and conditions of access (e.g. rights, charging);

    (b) it is made available under some open access (free at the point of use) arrangement;

Recent work has concentrated on repositories of scholarly publications, accompanied with a trend towards institutional repositories rather than thematic, centralised, repositories. However the associated metadata conventionally is DC (Dublin Core) [DC] and the interoperation protocol is

DCwrapped with OAI-PMH [OAI] and distributed query/retrieval is accomplished by OAISTER.

This architecture has a relatively informal syntax and no appropriate mechanism for defined semantics although in 2007 some improvements were made by projects trying to recast DC in RDF [Po07]. This architecture is not suitable for other repositories (e.g. research datasets, software) and does not integrate with the formal syntax and defined semantics of CERIF (Common European research Information Format) [CERIF] CRIS (Current Research Information Systems).

2. Benefits of Research Management Information

The benefits of research management information include:
1. To allow a research organisation to have an inventory of its research activity;
2. To allow evaluation of research output value, impact (RAE (Research Assessment Exercise in UK) type activities);
3. To allow a research organisation to make strategic decisions;
4. To allow innovators / entrepreneurs to find research ideas to take to wealth creation or improvement in the quality of life;
5. To allow the public to appreciate the research.

Research management information encourages improvements in the research process, notably:
1. Faster research turnround (communication of results, stimulation of new work to confirm or disprove) and therefore more and faster progress;
2. Originator improved quality – because of access & review by peers causing improvement in the research performance of the individual;
3. Community improved quality – because of access & review and discussion by peers raising the shared knowledge of the community;
4. Improved innovation and communication to innovators / entrepreneurs to stimulate wealth creation;
5. Improved education by faster provision of the latest research results to educators;
6. Improved public engagement through facilitation of access by the media to write – from the available information – research 'stories';
7. Improved PR for the institution through demonstration of their research output in quantity and quality of publications, patents and products, listing of research funding obtained, cataloguing expertise of research staff and advertising capabilities with research facilities and equipment;

From this we may conclude that each university and research funding organisation should have some sort of CRIS which can be used to track persons, organisational units (eg departments), research projects and their funding and also research outputs (publications, patents, products). Upon these research outputs performance may well be judged with financial implications. In

addition, at each institution there should be a repository of scholarly publications and a repository of research datasets and software.

3. Repositories

Repositories provide a digital store of research outputs. The usual kind of repository stored scholarly works or papers. Commonly these repositories store white literature. This is peer reviewed material – usually papers in journals or conference proceedings and monographs. Ideally the material is deposited – with associated metadata - in the repository upon acceptance for publication and is the final version as sent to the publisher post peer review. Some repositories store a copy downloaded from the repository of the publisher. Some store in the their repository the DOI (Digital Object Identifier) of the version in the publisher's repository. Others store also earlier versions of the material; however it is necessary to identify clearly the peer-reviewed version. Some few publishers demand an embargo period before access is open; nonetheless the metadata will indicate that the material is available, peer-reviewed and published and so any researcher can request a copy from the author under 'fair use' principles. The Eprints repository software [ePrints] even provides a 'button' for this purpose.

Many repositories store, additionally, grey literature. This is material that may – or may not – have been peer-reviewed but where the publisher does not have as their primary objective commercial publication. Grey literature includes internal technical reports, management information and newsletters and ephemera (such as brochures describing products or services). Grey literature may be made available open access although there may be restrictions on access, especially since some material (e.g. technical reports) may include valuable intellectual property of the institution.

Repositories may be institutional or thematic. The advantages of institutional repositories include management control of the deposit process; production of an inventory of research output and increased visibility and status for the institution. The advantage of thematic repositories is bringing together in one place – possibly with consistent metadata – articles relating to one subject domain. The best known is arXiv [arXiv]. Thematic repositories can be produced on demand by harvesting across institutional repositories.

Repositories are searchable from within and outside the organisation. Usually provided is a simple search (like Google) on the metadata (equivalent to a library catalogue card); the retrieved list of items can be browsed to determine which are of real interest and then access to the full text (or multimedia) article can be obtained. Some repositories provide full text (or multimedia) search on the article itself although this requires more computing resource.

Clearly the advantages of a repository of scholarly material to an institution are: to record the research output of the organisation, to demonstrate the scholarly output of the organisation to the world, to achieve a faster research 'cycle' by

early availability – and therefore more research discoveries and improve citation counts on the material from the institution caused by open availability.

Fortunately, the vast majority of commercial publishers do not object to researchers depositing in their own repository – see SHERPA [Sherpa] for details of publisher policies on deposit.

Access is available to:
1. anyone using a web browser and knowing the URL (address) of the repository;
2. Or having it stored in their favourites/bookmarks associated with their web browser;
3. anyone searching one repository which is linked to all others - a protocol called OAI-PMH does this;
4. anyone searching a CRIS linked to others and to local repositories - which provides also research context;
5. via Google, Google Scholar etc which indexes - by harvesting - over all repositories made available and provides search on this index (metadata).

4. CRIS (Current Research Information Systems)

The concept of CRIS has been around for many years. There is a clear advantage to a research institution or a research funding organisation to have management information on its research activities. Most research funding organisations had systems available in the seventies. In the eighties initiatives were started to interoperate these CRIS to share information among different funding organisations or research institutions. The IDEAS project 1984-87 and the succeeding EXIRPTS Project 1987-1989 demonstrated feasibility; the EC (European Commission) funded a group of national experts 1989-1990 to produce the CERIF recommendation to EU member states [CERIF91]. Experience in the nineties led to the group being reconvened in 1997 to produced the much better CERIF2000 recommendation in 1999. The EC requested euroCRIS (www.eurocris.org) to curate and develop CERIF. Many CRIS are CERIF-compliant to some degree, ranging from ability to interoperate using CERIF to full compatibility with the latest (CERIF2006) version.

The CERIF structure (simplified) is represented in **Figure 1** ; the major features are:
1. a formal syntax (structure) which is extensible from the mandatory core;
2. defined semantics and the ability to utilise any set of declared semantics;
3. built-in mutlilinguality with Unicode character representation and language variants of all text fields;
4. the use of linking relations with role, date/time start, date/time end between all major entities. This provides great flexibility in

representing a fully connected graph of relationships (not limited to a hierarchy) and to express accurately relationships;

5. linking relations can also be related to any one entity allowing recursion – this is particularly useful in relating instances of organisational units or persons;

Error! Objects cannot be created from editing field codes.
Figure 1: CERIF Model (simplified)

It is to be noted that Result Publication is an entity within the model and is linked to all the others thus providing a much richer context than that provided by a conventional repository [AsJe04] and is closer to the IFLA model [IFLA]. CERIF has great expressive power **Figure 2**; **Figure 3**:

Error! Objects cannot be created from editing field codes.
Figure 2: CERIF Expressivity

Error! Objects cannot be created from editing field codes.

Figure 3: Graph of CERIF Expressivity
In many implementations a CERIF-CRIS has been linked to a repository of scholarly articles. The formal metadata in the CRIS provides much more accurate (relevance) and complete (recall) retrieval and additionally relates the scholarly article to its research context including persons, organisations, projects, facilities, equipment, events etc. Examples of CRIS and repository linkages are in **Table 1**.

Country	Institution	Repository	CRIS
NL		DARE	METIS
NO	UiB	DSpace	FRIDA
SL		COBISS	SICRIS
GB	STFC	ePubs	CDR

Table 1: CRIS and Repositories

5. Bringing it all together

Linkage between repositories is commonly accomplished using the OAI-PMH protocol wrapping DC metadata and OAISTER search. This – as has been indicated – has difficulties because the DC metadata is insufficiently formal in syntax and has ambiguous semantics.

However, linkage can be achieved by attaching to the repository a CRIS (which the organisation should have in existence anyway for its management purposes) **Figure 4**. Then the full power of CERIF can be used for interoperability **Figure 5**.

Of course, if the repository is opened to harvesting then it can also be accessed by Google and/or Google Scholar providing another route for access.

The key is formalised metadata [Je00].

Error! Objects cannot be created from editing field codes.
Figure 4: CRIS & Repositories at one institution

Error! Objects cannot be created from editing field codes.
Figure 5: CRIS and Repositories Interoperating

6. Conclusion

Each research institution should have a repository of research output publications: repositories are available, tried and tested, there is freeware and lots of enthusiastic assistance, IT staff can implement a repository system relatively easily, repositories demonstrate benefits in research uptake, citations and assessment scores.

Moreover, the repository should be linked to the existing or planned CRIS of the research institution to have full research context and improved access, accuracy (precision or relevance), completeness (recall) in searching and interoperability.

Research institutions should be prepared for the new e-research environment: this involves implementing a CRIS and repositories.

There is a lesson to be learned from the experience in Norway, and their strategy is being taken up progressively in other European countries. There the funding to a university in year n+1 is dependent on performance in year n. Performance is measured by research publications, grant income and PhD production. Publications (and research income) are recorded in a CRIS (FRIDA) + repositories. We should ask ourselves how long until this is Europe-wide?

References and Bibliography

[arXiv] http://arxiv.org/

[AsJe04] Asserson, A; Jeffery, K.G.; 'Research Output Publications and CRIS' in A Nase, G van Grootel (Eds) Proceedings CRIS2004 Conference, Leuven University Press ISBN 90 5867 3839 May 2004 pp 29-40

[CERIF91] van Woensel, L: (1988) CERIF Manual October 1988

[CERIF] www.eurocris.org/cerif

[DC] http://purl.oclc.org/metadata/dublin_core/

[DSpace] http://www.dspace.org/

[ePrints] http://www.eprints.org/
[ePubs] http://epubs.cclrc.ac.uk/

[FRIDA] http://frida.uio.no

[IFLA] http://gopher.konbib.nl/dutchess/06/71/info-3501.html

[Je99] Jeffery, K G: 'An Architecture for Grey Literature in a R&D Context' Proceedings GL'99 (Grey Literature) Conference Washington DC October 1999

[Je00] Jeffery, K G: 'Metadata': in Brinkkemper,J; Lindencrona,E; Solvberg,A (Eds): 'Information Systems Engineering' Springer Verlag, London 2000. ISBN 1-85233-317-0.

[Je05] K G Jeffery: 'CRIS + Open Access = The Route to Research Knowledge on the GRID' Invited talk; IFLA2005 Oslo August 2005; Conference Proceedings Session 101 (in English, French, Russian)

[Je07] Keith G Jeffery 'Infrastructure and Policy Framework for Maximising the Benefits from Research Output' Keynote in Leslie Chan and Bob Martens (Eds) Proceedings 11th International Conference on Electronic Publishing (ELPUB2007) 'Awareness, Discovery and Open Access' Vienna Austria 13-15 June 2007 pp 1-12 IRIS-ISIS Publications 2007 ISBN 978-3-85437-292-9

[JeAs06] K G Jeffery, A G S Asserson 'Grey in the R&D Process'; The Grey Journal Vol 2 Number 3 September 2006 ISSN 1574-1796

[MARC] http://minos.bl.uk/services/bsds/nbs/marc/commarcm.html

[OAI] www.openarchives.org

[Sherpa] http://www.sherpa.ac.uk/

[Po07] http://www.ukoln.ac.uk/metadata/dcmi/rdf-values/

[W3C] www.w3.org

[W3Cmetadata] http://www.w3.org/Metadata/