New Review of Information Networking

ISSN: 1361-4576 (Print) 1740-7869 (Online) Journal homepage: https://www.tandfonline.com/loi/rinn20

Institutional Repositories and Current Research Information Systems

Keith Jeffery & Anne Asserson

To cite this article: Keith Jeffery & Anne Asserson (2009) Institutional Repositories and Current Research Information Systems, New Review of Information Networking, 14:2, 71-83, DOI: 10.1080/13614570903359357

To link to this article: https://doi.org/10.1080/13614570903359357

New Review of Information Networking, 14:71–83, 2008

ISSN: 1361-4576 print / 1740-7869 online
DOI: 10.1080/13614570903359357

Routledge
Taylor & Francis Group

INSTITUTIONAL REPOSITORIES AND CURRENT RESEARCH INFORMATION SYSTEMS

KEITH JEFFERY

Science and Technology Research Council, Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Oxfordshire, UK

ANNE ASSERSON

University Library, University of Bergen, Bergen, Norway

IRs (Institutional repositories) with deposit by the author of the 'green' peer-reviewed publication provide—through OA (open access)—improved access and intellectual property inventory. Increasingly organizations and research funders mandate deposit OA preferably in an IR. Publishers offer OA by author payment. CRIS (Current Research Information Systems) cover the research activity of an organization. CERIF (Common-European Research Information Format) is an EU recommendation to member states for CRIS. CERIF allows interoperability across CRIS. CERIF provides metadata describing publications with formal syntax and declared semantics. The CRIS provides the research context for the publication and links to associated research datasets and software.

Keywords: *integration, metadata, repositories, research information*

The Requirement

Introduction

The given wisdom is that maximizing access to research results improves wealth creation and the quality of life, the pace of research, and the quality of research. Much research is publicly funded and therefore access to the results of that research should be open and free. Such access has implications in technical, legalistic, management, and economic dimensions. The hypothesis of

Address correspondence to Keith Jeffery, Science and Technology Research Council, Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Chilton, Didcot, Oxfordshire OX11 0QX UK. E-mail: keith.jeffery@stfc.ac.uk or Anne Asserson, University Library, University of Bergen, Nygardsgaten 5, Bergen, Norway. E-mail: anne.asserson@fa.uib.no

this paper is that access to research publications is facilitated by a CERIF-CRIS and, furthermore, that the access is enriched by the contextual information in the CRIS and the access provided via the CRIS to further relevant information.

THE USER GROUPS

The researcher requires access to find relevant pre-existing research output and to find possible research collaborators. The research manager requires access to check completeness of recorded outputs from her institution, to compare with that of other institutions and thus to develop strategy for her institution. The funding agency requires access to ensure defined outputs from the funded research proposal are delivered, to compare outputs with those from other funding agencies, and to find appropriate referees. The policymaker requires access to compare outputs produced by different continents, countries, institutions, and research teams. The innovator requires access to find new ideas which are exploitable for wealth creation or improvement in the quality of life. The educator requires access to obtain teaching material. The student requires access to use learning material. The media require access to obtain information that can be recast as "stories" which popularize research or raise social, ethical, political, or economic issues concerning the research for the public interest.

THE ACTIVITY

The provision of more complete and accessible results of previous research improves the review of previous research before commencing a new research project. Wasted effort will be avoided and a better (novel) idea will be formulated. Discovering an applicable and appropriate technique—such as an experimental protocol, or a computer program for simulation or statistical reduction—from another domain in cross-disciplinary research can be valuable and stimulating. Furthermore, as a by-product, a researcher may find a potential collaborator or complementary co-worker for a research idea.

Increasingly, the research performance of an individual, a group, a department, faculty, or university is evaluated based on research output. The more complete and accessible outputs are, the better the quality of the evaluation. The metrics imposed on

the raw data (i.e., how one ranks different publication channels such as journals or uses online accesses and downloads or count citations) are a separate issue which can be (and has been) debated energetically. However, without complete and verifiable raw data, evaluations are worthless. Data quality is improved if the data have formal syntax (structure) and declared semantics (meaning). This allows for improved data collection with constraints on allowable values and suggested values, validation against accepted values, and improved utilization with query improvement and result explanation.

Utilization of scholarly publications has many aspects. Summary information may inform strategic decisions on research funding or areas of priority in a research institution. The publication, itself, provides a source of ideas and demonstrates their potential use. This may be used by the entrepreneur or innovator who wishes to invest venture capital to create products or services with associated wealth creation (jobs, profits for shareholders), and the provision of contextual research information from a CRIS greatly improves the utility for entrepreneurs.

Past research output, in the form of publications, forms the basis of today's teaching material. With the increasing volume of material and the pace of change, educators need easy access to research material to improve their curricula. The students also benefit; with increasing self-learning even in structured educational environments, the students need easy access to research publications as well as learning materials.

The public require access to research information. This is usually provided in a digestible form via the media who popularize science with appropriate "stories." It is best for everyone that such stories should be based on trustworthy scientific results and that the journalists should be able to access these easily and openly.

CONCLUSION

From the above it is clear that a range of end-users require easy (fast, efficient) access to research output material and its presentation in an understandable form. Technically, the implication is for high quality descriptive metadata, fast searching of metadata, fast searching of text and multimedia, and well-structured results—with contextual information to inform the requested

information (e.g., a publication). Research information is—by its nature—distributed among various organizations with differing roles such as funder or research institution. For the end-user access to heterogeneous distributed CRIS and repositories should appear homogeneous and local to the end-user. From decades of research it is well-known that instead of reconciling each source to every other source, it is much more efficient to reconcile to a canonical syntax (structure) and semantics (meaning). This involves translation of character sets, language, and ontological terms. Legalistically—while the requirement is for toll-free open access, restrictive metadata may document for software to enforce—claimed rights which should be respected (like attribution) and may even define a price for access. Economically, there is a need for a business model where costs of production and utilization of research output material lie where they fall but the intermediate access is free. For most purposes, the end-users require the research output material to be presented within the context of the research project, researchers, organizations involved, facilities and equipment, funding, etc.

CERIF-CRIS

The development of Current Research Information Systems (CRIS) has a 40 years history. Currently an EU Recommendation to member states, Common European Research Information Format (CERIF) is being adopted quite widely and it encourages interoperation. A CRIS typically has information on projects, persons, organizational units, funding programs, research outputs (products, patents, and publications), facilities and equipment, and events. The novelty of CERIF is:

- its formal data structure;
- its use of linking relations to allow n:m relationships with declared role and temporal duration;
- its use of multiple character sets; and,
- its provision of multilinguality.

The formal data structure ensures data integrity and avoids multiple instances of the same attribute values. It also makes for efficient data processing. The linking relations permit the representation of

a fully-connected graph that represents the real world much more accurately than simple data structures such as hierarchies. The role attribute allows great expressiveness and—with declared semantics—allows not only clear understanding of the information but also mapping across heterogeneous sources using domain ontologies or thesauri. With UniCode as a standard multiple character, sets can be represented supporting multilinguality. In CERIF, all text fields can be represented in any language with appropriate attribution of the language and the kind of translation— again this allows for interoperation across heterogeneity.

To illustrate the linking relations concept, consider the following case illustrated in Figure 1: A person A is an employee of organization O and a member of organizations M and N, both of which are parts of O. She is author of X in which O claims the IPR (intellectual property right) and project leader of P. In CERIF the following records would be in base tables: Person: A; OrgUnit: O,M,N; Publication: X; Project: P. The link tables would be: Person-OrgUnit: A-employee-O, A-member-M, A-member-N; OrgUnit-OrgUnit: M-partof-O; N-partof-O; Person-Publication: A-author-X; OrgUnit-Publication: O-IPR-X; Person-Project: A-projectleader-P. In fact, the link tables include, as well as role, the temporal information concerning start and end date-time. In this example, it may be that when A authored X she was no longer a member of M. This relatively simple example illustrates the power of CERIF as a data model.

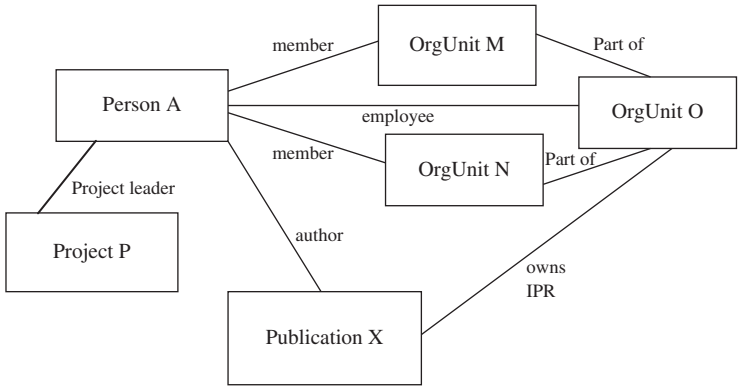

FIGURE 1 Example of CERIF.

CERIF is maintained by the not-for-profit organization euroCRIS (www.eurocris.org) from whence details are available. Commercial CRIS offerings are available from uniCRIS at www.unicris.com which is fully CERIF-compatible, Atira (PURE System) at www.atira.com, and Avedas at www.avedas.com. Many funding agencies and research institutions have some form of "home-brew" CRIS; the majority are more-or-less CERIF-compatible. The provision of CRIS in a modern e-infrastructure environment has been discussed in Jeffery (2004).

Repositories

Repositories store and provide access to the full text (or multimedia) of the scholarly publication. Although there have been some attempts to also use a publications repository for research datasets, it is usual to separate the publication repository from the e-Science (e-Research) repository of research datasets and software. This is because of their different access patterns and different metadata requirements. The e-Research repositories require much more detailed metadata to facilitate and control utilization of the software and datasets, in addition to the simpler metadata to allow discovery of the resources. Most e-Research repositories today are usually specific to an individual organization and built using "homebrew" software because of their novelty and the differing requirements on metadata imposed by different (commonly international) communities, e.g., in space science, atmospheric physics, materials science, particle physics, humanities, or social science. Publication repositories typically use some form of Dublin Core Metadata (DC) (see http://dublincore.org/), and most are Open Archive Initiative – Protocol for Metadata Harvesting (OAI-PMH) compliant for interoperation and are indexed by Google Scholar (see www.openarchives.org/OAI). Example software systems are www.eprints.org, www.dspace.org, www.fedora.info, and epubs.cclrc.ac.uk.

Metadata, Access, and Interoperation

Digitally-created articles rely heavily on both the metadata record—to support fast, easy access—and the articles themselves—to allow full text or multimedia searching—being deposited. International

metadata standards and protocols must be applied to repositories so that retrieval may be consistent with appropriate recall (precision) and relevance across heterogeneous repositories. A model for formalizing metadata is required and has been suggested (Jeffery 2000).

Current interoperable repository technology is using OAI-PMH with DC or even Object Re-use and Exchange (ORE) (see www.openarchives.org) as packaged metadata. Examples of such an approach have been utilized in the DELOS project and NoE (see www.delos.info) and DRIVER followed by DRIVER II (www.driverrepository.eu). However, it is the experience of the authors that this is insufficient to meet the requirement for repositories and certainly insufficient to provide the interoperation of contextual metadata in CRIS. The DC 13- then 15-element metadata standard does not have a sufficiently formal syntax, nor declared semantics, for effective interoperation processing. Although DC has been extended (Qualified DC) to improve the situation and recent work has extended DC with domains and ranges and even produced a Resource Description Framework (RDF) representation this does not fully overcome the problem. This may be characterized as the need for machine-understandability as well as machine-readability of the metadata. At present,, interoperation of repositories depends on the end-user reading the metadata, understanding, and choosing which items represented by the metadata to access. This is time-consuming and does not scale. Furthermore, the research output should be understood in context—that is the publication or research dataset related to the research projects, persons and their roles, organizational units, funding, research facilities and equipment, etc., involved in the research which generated the output. One example should suffice to explain the difficulty using DC. The element contributor is defined at http://purl.org/dc/elements/1.1/contributor as:

> Label: Contributor
> Definition: An entity responsible for making contributions to the resource.
> Comment: Examples of a Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity.

The example illustrates, exactly, the problem: the "type" of the element contributor is not defined (although with namespaces, domains, and ranges, a limited set of acceptable lexical terms can be defined). The kinds of contributor in the example would likely have a different legal status, rights, and responsibilities but there is no syntax nor semantics for recording this. There is no concept of the relationship between contributors (except that there is, confusingly, another element named creator (and the definition has a comment or description exactly as for contributor). Since 1999 (early in the life of DC) these criticisms have been made by members of euroCRIS and the alternative approach based on formal syntax and defined semantics (described in the following section) proposed.

In fact, the DRIVER consortium, itself in a public paper, http://www.driversupport.eu/documents/DRIVER_ Review_of_ Technical_Standards.pdf criticizes DC as unsuitable, criticizes other formats (such as Metadata Object Description Schema (MODS)) and even mentions CRIS and CERIF. A later (and very recent) paper http://www.driversupport.eu/documents/D4%203_ Tech_Watch.pdf is more specific on the need for CERIF-like metadata and lifts much information from the euroCRIS website in section 2.3. Despite this later paper recommending integration of CRIS and OA repositories, there is no technical proposal of how this should be done.

Similarly the Knowledge Exchange http://www.knowledge-exchange.info/ (which has an intersection of members with DRIVER) has considered the relationship between CRIS and repositories and initiated a project (2008–2010) on this. This project was instigated following a meeting where a euroCRIS member presented the case for integrating CRIS and repositories. Both DRIVER and Knowledge Exchange have claimed there is no method for interoperation between a CRIS and OA repository; in fact, within the euroCRIS community there are several working examples (for example in UK, Norway, Flanders, and Denmark as mentioned in section 5), and the following solution proposed is based on the euroCRIS members' experience of this. In the later DRIVER paper http://www.driversupport.eu/documents/ D4%203_Tech_Watch.pdf, some case studies indicate, in overview, that such a linkage is possible thus contradicting the earlier statement.

To summarize the view from the euroCRIS community: the current DC metadata standards and OAI-PMH for interoperability are insufficient for scalable, automated retrieval with appropriate relevance (precision) and recall. Current DC is machine-readable but not machine-understandable. The underlying problem is that a formalized syntax and semantics (vocabulary) for each relevant DC element was not specified in "simple DC." This has partially been remedied by the use of namespaces in "qualified DC" as illustrated previously. A second problem concerns the element set tags "contributor," "creator," and "publisher," which are actually in the real world (and as mapped in CERIF) roles of a person or organizational unit and should be represented by a relationship (between the article and the person or organizational unit) where the role belongs to a namespace and is temporally limited in the way CERIF represents this situation. A third problem is the tag "relation" which is extremely general; the real world is much better modeled through typed relations with role and temporal validity. Other problems include the tag "coverage" which only recently has been separated into temporal and spatial aspects; yet, these are fundamental retrieval criteria for much material. A formalized version of DC overcoming these limitations has been suggested (Jeffery 1999) and defined (Asserson and Jeffery 2005) to also form part of the CERIF model allowing tight integration with CRIS. Recently, the DC community has recognized these problems and, with more recent work available at http://dublincore.org/documents/abstract-model/ and http://dublincore.org/documents/dc-rdf/ is attempting to address them.

There is one final remaining issue: preservation and curation of research output. There is current work (OAIS) to define metadata standards to achieve this (available at http://ssdoo.gsfc.nasa.gov/nost/isoas/), but this is really only a proposed architecture. Major problems are concerned with maintaining the articles on current (i.e., usable) media—which implies regular media migration—and maintaining, alongside research datasets, appropriate software with the environment in which that software executes. This whole subject area is fraught with difficulties in technical, management, legal and economics dimensions. It relates to records management and thus to "freedom of information" and "data protection" or "privacy protection" legislation which varies by country. The technical problems of preserving a software environment to ensure

access, or migrating software periodically to a new environment with assurance of equivalent functional operation are immense, with concomitant costs.

Linking Institutional Repositories and CRIS

An architecture for providing a complete research information environment at an institution is presented. The linking together, at an institution, of a "green" OA repository of articles (that is a repository of publications deposited institutionally for toll-free open access in parallel with a peer-reviewed publication), a CRIS (to provide contextual information), and an OA repository of research datasets and software (Jeffery and Asserson 2006a). (Figure 2) provides that institution with an information resource suitable for all the end-users and roles discussed earlier. Furthermore, the formalized structure of the CRIS allows a reliable workflow to be engineered which, in turn, encourages deposit of research outputs by reducing the effort threshold by using intelligent prompts or suggestions based on the information already stored and any constraints on permissible values of attributes. Such a system is being implemented progressively at STFC Rutherford Appleton Laboratory by the e-Science team where the CERIF-CRIS is named the Corporate Data Repository, the OA repository is ePubs and the e-research repository is the e-Science repository. Similar linking of

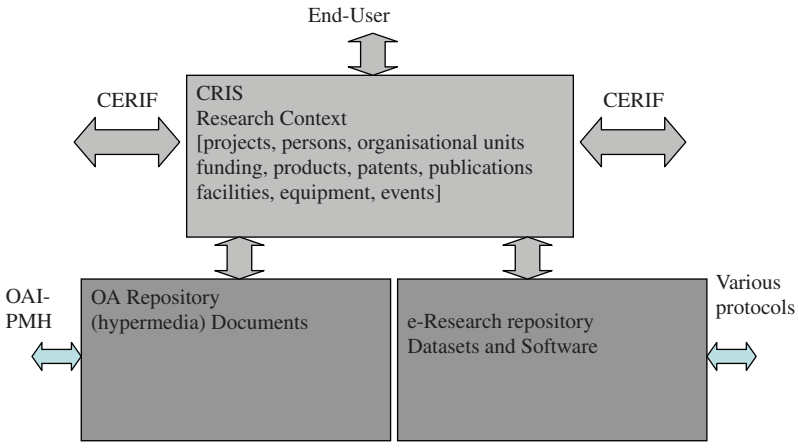

FIGURE 2 Architecture for an institution.

CRIS and repositories is underway, for example, in Norway (FRIDA to NORA), Flanders (FRIS) and Denmark (PURE).

However, the requirements of the end-user extend beyond the individual research institution or funding organization. The institutional CERIF-CRIS system can be linked to others because they have a formal structure and, hence, can be interoperated reliably and in a scalable way (Jeffery 2005). This, in turn, provides a network of access to institutional OA repositories (of articles) or e-research repositories linked to each institutional CRIS via the CERIF-CRIS gateways, enhancing and controlling the access using the CERIF-CRIS information as formalized, structured, and contextual metadata which is more detailed than DC and suitable for intelligent (machine-understandable) interoperation (Figure 3). Successful interoperation of CERIF-CRIS has been demonstrated, including for euroHORCS (European Heads of Research Councils) in October 2006. However, as yet, the whole architecture has not been demonstrated although such a demonstration is being planned.

The key point is that the metadata for a publication (or dataset) is stored in the CERIF-CRIS with formal syntax and with defined semantics and the repository just acts as a deposit space. In this way, management information and analysis can be done using the (formal) CERIF-CRIS, while retrieval of the individual publication (or dataset) is done through the repository sanctioned by the CERIF-CRIS. The repository may or may not also store metadata (usually in DC, for OAI-PMH interoperation and OAISTER (http://www.oclc.org/oaister/) retrieval) but this metadata is best generated from the CERIF-CRIS. This is because the CRIS, in a research institution, is intimately linked to the researcher workbench and organizational workflow, and much of

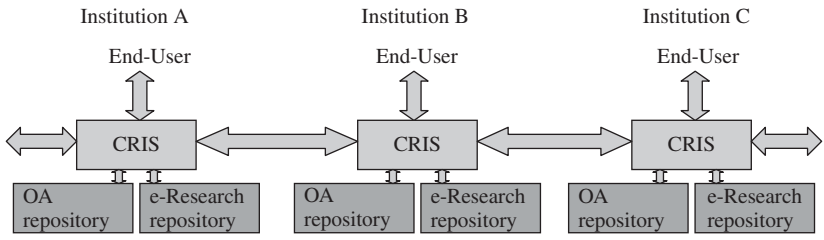

FIGURE 3 Architecture for OA.

the metadata required for a publication (author, institution, rights) is already stored in the CRIS and does not need the author to re-input. Furthermore, the publication metadata is surrounded by relevant contextual information of use to the end-user.

This proposal is not in-line with mainstream thinking. Current practice—especially among librarians—is to utilize whatever metadata is available in the repository system—commonly heterogeneous, with rather informal syntax and semantics—and interoperate using DC. The proposal here is that there is better quality metadata in a CERIF-CRIS which has advantages—formal syntax and declared semantics—both for retrieval and for interoperation.

Looking Forward

Let us speculate on a possible future if, and when, the architecture described previously is implemented. Let us assume the OA publication repository and research dataset and software repository are linked together by and accessed through a CERIF-CRIS at each institution. One might change the business model and workflow of scholarly publication. The author deposits in an open access "green" repository (technically this is a submission for publication) so instead of submitting in parallel to a journal or conference peer-review process, the peer-review is done either by:

(a) a learned society managing a "college" of experts and the reviewing process—for a fee paid by the institution of the author or by the author;
(b) allowing annotation by any reader (with digital signature to ensure identification / authentication);

in both cases being alerted by "push technology" that a new article matching their interest profile has been deposited.

The former peer-review mechanism would maintain learned societies in business, would still cost the institution of the author or the author, but would probably be less expensive than publisher subscriptions or "gold" (author or author institution pays) open access. The latter is much more adventurous and in the spirit of the internet; in a charming way, it somehow recaptures the scholarly process of two centuries ago (initial draft, open

discussion, revision, and publication) in a modern world context. Either scheme separates peer review from publishing and publishing from access and utilizes different business models for each stage. Certainly, such schemes would considerably reduce the costs of research publication, would increase widespread dissemination, would encourage greater participation in the research process, and increase the funding available to research because of the reduced costs.

References

Asserson, A. and K. G. Jeffery. "Research Output Publications and CRIS" *The Grey Journal* 1.1 (Spring 2005):5–8. TextRelease/Greynet ISSN 1574–1796. Print.

Jeffery, K. G. "An Architecture for Grey Literature in a R&D Context." *Proceedings GL'99 (Grey Literature) Conference*. Washington, DC: October 1999. Web. Retrieved in 2000 from http://www.konbib.nl/greynet.

Jeffery, K. G. "Metadata" *Information Systems Engineering* Eds. J. Brinkkemper, E. Lindencrona, A. Solvberg. London: Springer Verlag. ISBN 1-85233-317-0, 2000. Print.

Jeffery, K. G. "The New Technologies: can CRISs Benefit" *Proceedings CRIS2004 Conference*. Eds. A. Nase and G. van Grootel (Eds), 77–88. Leuven, Belgium: Leuven University Press. ISBN 90 5867 3839. May 2004. Print.

Jeffery, K. G. "CRISs, Architectures and CERIF CCLRC-RAL" *Technical Report*. RAL-TR-2005-003. 2005. Print.

Jeffery, K. G. and A. Asserson: "CRIS Central Relating Information System" "Enabling Interaction and Quality: Beyond he Hanseatic League. Eds.in Anne Gams Steine Asserson and Eduard J Simons": 109–201. *Proceedings 8th International Conference on Current Research Information Systems CRIS2006 Conference, Bergen*, May 2006. Leuven, Belgium: Leuven University Press ISBN 978 90 5867 536 1. Print.