Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2014) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

CRIS 2014

Research data meets research information management: Two case studies using (a) Pure CERIF-CRIS and (b) EPrints repository platform with CERIF extensions

Anna Clements[a], Valerie McCutcheon[b]

[a]*University of St Andrews, UK*
[b] *University of Glasgow, UK*

Abstract

This paper will describe how two research-intensive universities in the UK, St Andrews and Glasgow, have worked together over several years and projects to develop their institutional research management systems to deliver services to support the rapidly evolving needs of funders, institutional policy makers and management, and, importantly, the researchers themselves. This challenge is particularly acute at the moment with 'Open Science' one of the hottest topics around with organisations and funders from the G8[1] downwards stressing the importance of open data in driving everything from global innovation through to more accountable governance; not to mention the more direct possibility that non-compliance could result in research grant income drying up. There is a need to work with those researchers that need support to develop research data management processes and infrastructures that complement their ways of working and not just impose box-ticking exercises. We will explain the strategies, systems developed, and concerns arising to date at our two Universities to help support researchers and managers in this (r)evolution.

Peer-review under responsibility of euroCRIS.

Keywords: CERIF; research information management; research data management; CRIS-IR; institutional repository

Peer-review under responsibility of euroCRIS.

1. Introduction

Despite using different technical infrastructures for our research information management we both face the same issues and have solved them in similar ways: through the pragmatic use of standards, continual dialogue with other stakeholders, particularly researchers and funders, and best practice information management principles.

2. Information Management Principles

This integrated model (whether CRIS-IR or IR with CERIF[2] extensions) is a prime example of the successful practical application of the principles of good information management:

- Data is entered once, as close to source as possible, and reused
- Data Stewards keep control of the data within their domain of expertise
- Data is available to those who need it and are authorised to access the data
- Data standards, such as CERIF and existing data sources, such as Web of Science and Scopus can be used

Fig.1 below is a stylised overview of the research information system at the University of St Andrews illustrating the reuse of data sources. We will discuss later in this paper how CERIF has played an important part in our thinking and implementation plans as well as, increasingly, the work that the CASRAI-UK[3] working groups cover in the development of common definitions and vocabularies to describe various aspects of the research landscape and processes.

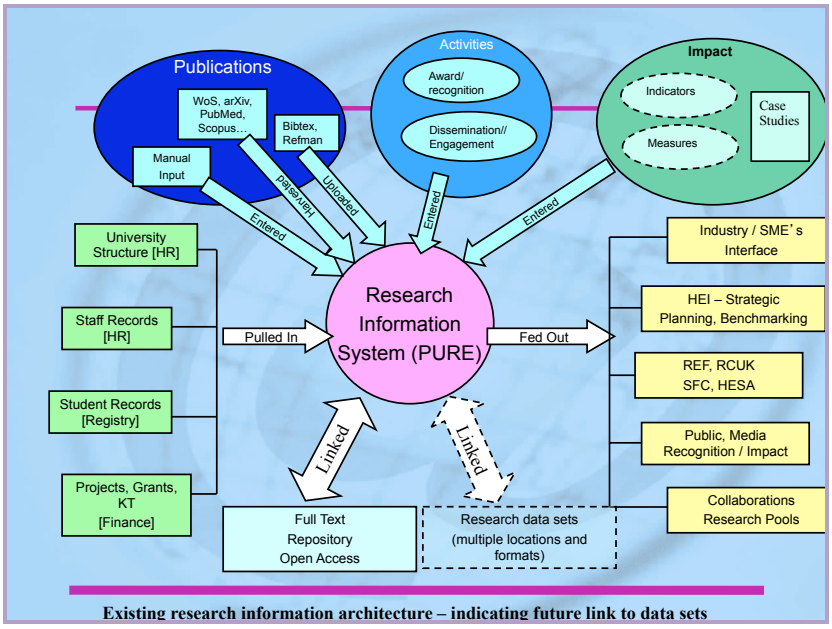

Fig. 1. Research Information System at University of St Andrews.

3. A brief history of Research Information at St Andrews and Glasgow

3.1. St Andrews

At the University of St Andrews, we have had an integrated research information infrastructure since 2006. The overall architecture has remained unchanged with a current research information system (CRIS) providing tools for managers and researchers to access all research-related institutional data from corporate systems such as human resources, student records, research grants and finance. In addition, the CRIS stores research outputs, outcomes, impacts and activities either via harvesting from third-party sources, such as Elsevier's Scopus [4]and Thomson Reuters' Web of Science[5], or via manual data entry by researchers.

The technology has been updated over the years, with an in-house CRIS being replaced by Pure[6] CERIF-CRIS from Atira (now part of Elsevier) in 2010. From the outset, the CRIS has been integrated with our open access (OA) institutional repository (IR) running in the DSpace[7] platform. The CRIS is the single 'golden' data source for the research publication metadata and, where a full-text version can be made OA, these metadata are pushed through to the institutional repository together with the full text. All workflow on copyright clearance and embargo periods is done in the CRIS. Thus the IR acts as a genuine repository of openly accessible documents.

This CRIS model has proved invaluable in promoting and improving research information quality whilst keeping burden to a minimum, whether via institutional driven processes such as the recent REF2014 (Research Excellence Framework)[8] national research assessment exercise, or academic-led initiatives to reuse information in individual and collaborative web sites.

3.2. Glasgow

At the University of Glasgow, we have had a Research System since 1994. The system holds information about research projects, applications and awards, and is linked to the Human Resources (HR), Finance, and Student Systems. In 2010 we added a regular upload of data from our Research System to our repository that holds publication and other research output details so that awards could be related to research outputs.

Like St Andrews these linked corporate systems are the authoritative source of data at our organisation and the data is re-used for many purposes including personal development and review, workload planning, reporting to funders and government (including the Research Excellence Framework exercises), and feeds to web pages. At Glasgow our repository 'Enlighten' [9]is the point of entry for metadata and full text associated with publications. Copyright, licensing, embargo periods, and metadata quality are managed by the repository team before any information is displayed publicly on the repository or re-used e.g. in web pages.

The University has undergone a number of system changes since 1994 including new finance, student and HR systems with a new Research System due for initial release early 2015. We believe that the strength of our information is in the specification and quality control of the metadata which can then be re-used regardless of the software that we utilise.

3.3. Working together

Both organisations have focused on improving the reuse and exchange of information both within the institution and externally. Our experience has been that this improves the quality of and trust in the data we hold centrally (e.g. in our Human Resources, Student and Finance systems) as well as decreasing the duplication of effort in collecting and validating the same data several times. There is still work to be done with external funders to develop better interoperability with our systems and this is another key area where we have been working together for several years to encourage dialogue between funders, institutions and system providers to implement effective data exchange procedures.

CERIF has played a key role is this work starting with the CRISPool[10] project in 2009, which brought heterogeneous data from St Andrews, Edinburgh and Glasgow together for the Scottish Universities Physics Alliance (SUPA); the IRIOS and IRIOS-2 [11]projects, that involved working with funders to "CERIFy" grant information and offer additional ways to link publications to grants; CERIF In Action (CiA)[12], produced a prototype CERIF-XML export/import function for reporting research outputs to funders and CERIF for Datasets (C4D)[13] examined the applicability of CERIF to recording metadata about datasets. Since 2012 we have been participating in the JISC supported CASRAI-UK working group pilot in the areas of data management planning, organisation identifiers, and open access & research outcomes. These working groups bring funders, institutions, suppliers and standards bodies together to discuss and agree common definitions and vocabularies – which are represented as use case application profiles - within the specific working group domain in order to facilitate data exchange.

We continue to work together and with other groups in the community. For example the EPrints User Group and the PURE User Group are exchanging ideas on further enhancements to their research data metadata specifications with a view to maintaining consistency and joining up discussions with all stakeholders.

4. So where does research data fit in?

4.1. Opportunities – extending the CRIS

As illustrated in Fig 1. extending the CRIS to include research data can be achieved by adding metadata to the CRIS with external links where appropriate e.g. DOI to the data itself. This mirrors the arrangements for the full text open access publications. The CERIF for Datasets project mentioned earlier, and the EU Engage project[14] project, have also successfully mapped CERIF to several common research metadata schemas, including CKAN, EGMS, DCAT[15] and MEDIN[16]. Both projects looked at a 3-layer model for metadata – discovery, usage (contextual) and domain-specific; CERIF is fully able to represent both the discovery and contextual metadata layers with some minor additions recommended by the CERIF for Datasets project[17]released in version 1.6 Summer 2012. Domain specific metadata has been explored by a number of initiatives. In many cases third party data repositories hold the specialist data and it is not necessary for research organisations to replicate this as long as the generic discovery metadata is available in our systems.

The first step we have taken is to develop our initial data registries as part of the Cerif for Datasets project after consulting across the community. The key fields included DataCite mandatory fields. We are now working with our respective User Groups (Pure and Eprints) to bring together best practice from different sites and ensure that the dataset modules in both solutions satisfy a wide range of requirements. Pure released version 4.18 Feb 2014 which promoted the dataset to a top-level entity and included additional metadata; a further release is planned for Oct 2014 which will add further metadata, integration with DataCite and improved workflow. Glasgow have had a live EPrints data registry since January 2013 with other EPrints sites have similar instances e.g. Essex 'ReCollect'[18] plugin was made generally available in March 2013. A new version of the standard module is imminent and will combine best practice from some of these instances as a result of EPrints User Group activity and further enhancements are expected as business requirements evolve.

Glasgow have also introduced some fields into the traditional repository (where publications metadata and full text reside) to capture information required by funders as to whether the research materials have been acknowledged in the paper. This related area is subject to development and standardisation pending the outcome of various initiatives working on defining metadata for open access including Vocabularies for Open Access (V40A)[19], the CASRAI-UK working group, and RCUK requirements. Looking further ahead we are exploring ways to capture data management plans so that metadata can re re-used e.g. to indicate internal storage requirements.

Fig. 2. Screen shot from Glasgow EPrints : research data metadata

In this way the CRIS can be used as the research data registry for the Institution and, just as with full-text articles, we can combine metadata-only records with metadata-plus-data records. The latter being for those research data sets for which there is no trusted external data repository available to preserve and share the data long-term and data therefore has to be stored in robust internal stores linked to the registry. Once in the CRIS, links to publications, funding and other research activities can be made to provide rich contextual information for use in compliance reporting to funders as well as for continually building up a web of activities, events and outputs to support impact case studies.

4.2. Barriers – differences to consider and how we have approached these

However, unlike full-text publications which are almost always stored as portable document format (PDF) within our IR with accompanying standard bibliographic metadata, research data sets are much more heterogeneous. The heterogeneity is evident from both the physical manifestation of the data (numbers and organisation of files, file formats and descriptive metadata) and by the (sub-)discipline or even instrument-specific procedures used to collect, analyse, combine and reuse the data. It is simple to reuse (read and understand) a pdf, assuming you are familiar with the language and the subject matter. For most data, even human-readable formats such as XML require considerably more expertise (or accompanying metadata and documentation) and often access to technology, such as proprietary software, in order to understand, reuse and/or verify.

Mitigating action can be taken by encouraging the use of open and standard data formats, but the biggest challenge arguably is in understanding just how much we as a central support team can and should provide services to researchers versus how much researchers should be supported in developing their research practice to manage research data appropriately. Fig. 3 illustrates the components of a research data management service.

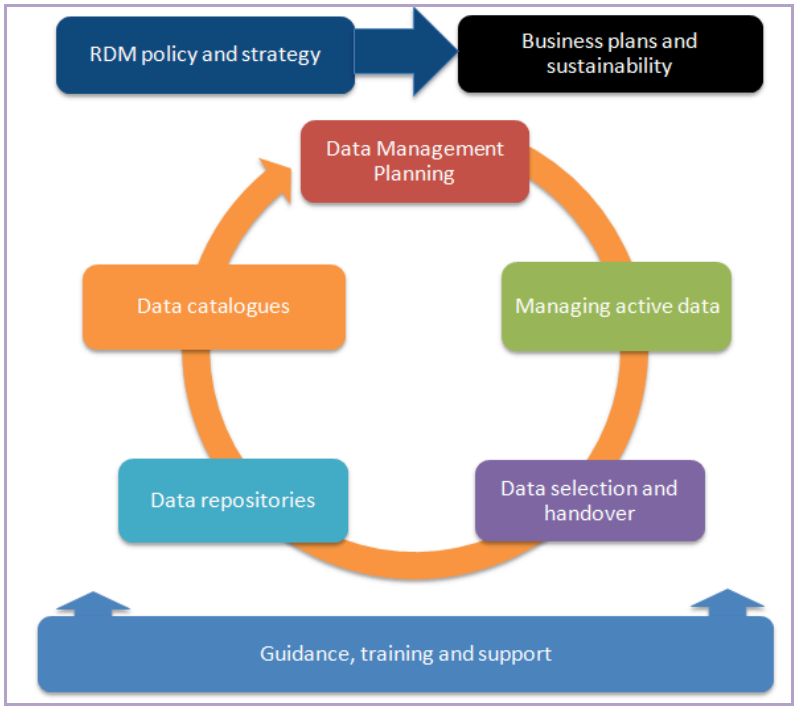

Fig. 3. Components of Research Data Management service, Digital Curation Centre[20]

Traditionally, repositories and catalogues have often been managed by the University Library and for data this model is repeated with both St Andrews and Glasgow University Libraries managing the extension of the CRIS-IR as data catalogue and institutional data repository (for those data that cannot find a home in an external data

repository or data centre). Data management planning, active data management and selection and handover tasks are either largely or entirely the responsibility of the researcher and in these areas considerable guidance, training and support is needed. Our approach is to work closely with research groups and administrators to identify and support individuals who can advise and train their colleagues in these areas – thus building a network of expertise across the institution. Given the subject-specificity of many of the processes, tools, data formats and research practice this seems a more sustainable approach than aiming to provide a heavily centralised service. Some traditional institutional repositories have encountered lack of engagement in gathering full-text. One reason might be that engagement is needed at an earlier stage of the publication process. The nature of full life-cycle research data management means that we have the opportunity to inform and support the researchers at a much earlier stage in the research process. The fact that many funders now ask for data management plan as part of the grant application process provides another driver for early engagement with the research process.

In addition, given the difference between disciplines in the type of data they work with and the processes carried out on the data there is a strong argument for encouraging researchers to deposit their data in an appropriate subject-specific repository. This is an area which is rapidly developing in terms of (inter)national infrastructure provision and should be encouraged by researchers, institutions and funders. The announcement in March 2014[21] of the merger of the re3data.org and Databib research data repository registries under the auspices of DataCite is an excellent move towards providing an authoritative list of trusted repositories.

5. Conclusion

Overall our CERIF-based institutional systems have proved to be flexible enough to cover discovery and contextual research data metadata and so fulfill the data registry requirement from our funders, as well as enrich our set of research information more generally. Thus research data has met up with research information in our existing infrastructure. It is though, important for all stakeholders to continue to work together to agree a common understanding of metadata about datasets so that standards, such as CERIF, can continue to be improved. This will facilitate current and future initiatives such as the UK national data registry and wider data registries in making information about datasets and the datasets themselves accessible to the widest possible audience.

Away from metadata standards, there is the considerably more complex task of understanding and engaging with researchers and their research processes to build up the network of data management experts who can bridge the gap between the policy, training and infrastructure support provided by the Library and Information Technology Services and the discipline-specific needs of the researchers. This is perhaps the biggest challenge that we face as we are effectively discovering and defining a new role (or roles) at the frontier between research and library services.

In the meantime we will continue to participate in a wide range of communications with internal and external stakeholders in order to further apply pragmatic solutions, building on or reusing our existing CRIS / CRIS-IR systems to support our researcher and institutional research information and data needs.

Acknowledgements

We would like to thank David McElroy for his work on the data registry at the University of Glasgow. We would also like to thank the members of the CERIF for Datasets project team[22] and Jisc who supporting this work [grant number DIINNAA].

References

[1]G8 Open Data Charter and Technical Annex
https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex (accessed 21 April 2014)
[2] Common European Research Information Format (CERIF)
http://www.eurocris.org/Index.php?page=CERIFreleases&t=1 (access 21 April 2014)
[3] The Consortia Advancing Standards in Research Administration Information (CASRAI), UK pilot project, JISC
http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/researchinformation/casraipilot.aspx (accessed 21 April 2014)
[4] Elsevier Scopus
http://www.elsevier.com/online-tools/scopus (accessed 21 April 2014)
[5] Thomson Reuteurs Web of Science
http://thomsonreuters.com/thomson-reuters-web-of-science/ (accessed 21 April 2014)
[6] Pure CRIS, Elsevier
http://www.elsevier.com/online-tools/research-intelligence/products-and-services/pure (accessed 21 April 2014)
[7] dSpace Repository software
http://www.dspace.org (accessed 21 April 2014)
[8] Research Excellence Framework 2014, Higher Education Funding Council England, Higher Education Funding Council Wales, Scottish Funding Council, Department for Employment and Learning Northern Ireland
http://www.ref.ac.uk (accessed 21 April 2014)
[9] http://eprints.gla.ac.uk/
[10] CRISPool project, JISC Research Information Management programme, 2010
http://crispool.org (accessed 21 April 2014)
[11] Integrated Research Input and Output System, IRIOS and IRIOS-2, 2011
http://irios2.wordpress.com (accessed 21 April 2014)
[12] Cerif in Action project, JISC RIM Programme 2011
http://cerifinaction.wordpress.com (accessed 21 April 2014)
[13] Cerif for Datasets project, JISC RIM Programme 2011-2013
http://cerif4datasets.wordpress.com (accessed 21 April 2014)
[14] EU Engage Open Data project
http://www.engagedata.eu (accessed 21 April 2014)
[15] Houssos N, Jorg B, Matthews B, A multi-level metadata approach for a Public Sector Information data infrastructure, CRIS2012 Prague Jun 6-9 2012
[16] Clements A et al, First workshop on Linking and Contextualising Publications and Datasets, Sep 26 2013, Valletta Malta
http://lcpd2013.research-infrastructures.eu (accessed 21 April 2014)
[17] Joerg B, Datasets in CERIF blog post, Cerifsupport.org
http://cerifsupport.org/2013/04/02/data-in-cerif/ and http://cerifsupport.org/2013/07/24/cerif-1-6-formal-models-released-for-testing/ (accessed 21 April 2014)
[18] Eprints ReCollect plugin
http://wiki.eprints.org/w/ReCollect (accessed April 23, 2014)
[19] Vocabularies for Open Access
http:// http://v4oa.net/about/ (accesses April 23, 2014)
[20]Jones, S, Pryor, G and Whyte, A, How to Develop Research Data Management Services - a guide for HEIs. In: DCC How-to Guides, 2013, Edinburgh, Digital Curation Centre:
[21]Datacite, re3data.org, and Databib Announce Collaboration
http://www.re3data.org/2014/03/datacite-re3data-org-databib-collaboration/ (accessed 21 April 2014)
[22] CERIF for Datasets Project Team
http://cerif4datasets.wordpress.com/c4d-team/ (accessed 21 April 2014)