JISC Research Information Management Programme: phase 1 technical synthesis

Document details

Author:	Rosemary Russell, UKOLN, University of Bath
Date:	4 May 2011
Version:	0.4
Notes:	Contact: <r.russell@ukoln.ac.uk>

UKOLN is funded by the Joint Information Systems Committee (JISC) of the Higher and Further Education Funding Councils, as well as by project funding from the JISC and the European Union. UKOLN also receives support from the University of Bath where it is based.

RIM1 technical synthesis

1 Introduction

The first phase of the JISC Research Information Management (RIM) programme funded five projects which ran from March until August 2010, representing a short timescale. The aim of this document is to provide a synthesis of the more 'technical' project outcomes; this will help to inform the second and later RIM phases, as well as providing a selective overview which may be relevant for other similar initiatives.

UKOLN also produced a briefing document[1] to support the JISC Call for RIM proposals in October 2010. This provides background on UK RIM activities as well as emerging themes. It is broader than the current document, including vocabulary alignment work and CERIF-based projects such as R4R which was part of the JISC Repository Start Up and Enhancement strand.

2 Scope

The synthesis is based mainly on the individual project final reports. 'Technical' is interpreted here in its broadest sense – the review attempted to find and highlight project activities relating to eg technical requirements, standards (CERIF in particular) and metadata, in addition to systems, software and infrastructure issues. Given the source material, the focus is more on technical *management* than the technologies themselves.

The report is inevitably more focused on some projects than others – two had a business/management approach (those led by UCL and Imperial College[1]), whereas the remaining three (BRIM, CRISPool, Enquire) included more technical elements eg stakeholder requirements analysis, metadata mapping, software implementation, evaluation. CRISPool was the most technically complete in that it comprised an end-to-end system – from user requirements, specification, mapping to CERIF, through to pilot implementation and use.

3 Thematic approach

Given the diversity of the projects it was difficult to identify common technical themes running across all five. However the following eight topics have been selected since they recurred frequently and/or were reported as significant; they are also important in the broader RIM context. The themes are described and illustrated in section 5 below.

- Lack of integration
- Research staff dissatisfaction
- Capturing RIM requirements
- Data quality
- Person identification
- CERIF
- Research outputs and impact
- Collaborative working

[1] The projects led by UCL and Imperial College have long titles with no acronym, so for ease of reference, the lead institution name is used instead throughout this report

4 Brief project context

In order to provide some context for the thematic sections, a very brief overview of each project follows, which mainly aims to highlight the more technical aims and outcomes, including systems implemented.

4.1 CRISPool: Using CERIF-XML to integrate heterogeneous research information from several institutions into a single portal

CRISPool[2] was led by the University of St Andrews. Its main objective was to build a portal exposing data (from heterogeneous, cross-institutional sources) on the web with basic search & retrieve functionality. It succeeded in using CERIF-XML to bring together data on people, organisations and publications from three universities for the SUPA (Scottish Universities Physics Alliance) research pool.

The project demonstrated that a dynamic searchable portal is much better than a fixed pdf publications list for SUPA reporting purposes. It is interesting that at least one other research pool has expressed interest in developing a similar portal.

4.2 Using Business Process Management Tools and Methods for Building Research Information Management (BRIM)

BRIM[3] was led by the University of Huddersfield. The project was aimed at investigating the feasibility of implementing RIM tools to satisfy the management needs of a university which is growing its research activity significantly.

RIM requirements were captured and a pilot system developed to demonstrate a research information layer which integrated with and harvested data from existing institutional systems in a standard, component-based way, and stored it in a model based on CERIF. Initial findings had indicated that it would be difficult to integrate and use data in a *fully* standardised, service oriented approach, so a slightly modified approach was adopted.

A further task investigated the use of ontologies for representing and storing publication information.

4.3 Enquire: Enrich and Research Outputs and Impact

The Enquire project[4] was led by the University of Glasgow. It aimed to widen the types of data captured for research outputs and impact at the University, in response to requirements including Research Council's UK (RCUK) Outcomes, the Research Excellent Framework (REF) and internal reporting.

The project built on the previous Enrich project that successfully linked the Institutional Repository (Enlighten) with other core University systems such as the Research System, meaning that outputs and impact can be linked to sources of funding that supported them.

One of the key findings from Enquire was that requirements specifications for output and impact reporting from research organisations to funders need further development. However the project has allowed the University of Glasgow to accelerate preparation of the environment for reporting.

4.4 Developing tools to inform the management of research and translating existing good practice

Project aims included developing an overview of the systems used by the institutions involved in the study and comparing the variety of tools available in the marketplace. 21 research-intensive institutions participated in the study and a range of staff involved in research were interviewed

The project identified 11 things that institutions want from research information. These information needs were mapped to available systems, revealing what functionality is provided. The matrix

produced demonstrated the fragmentation of provision over the range of needs – most suppliers only address a few requirements. The project was led by Imperial College London[5].

4.5 Defining a new role: the embedded Research Information Manager

This project was the least relevant to the current report, given its focus on the role of a Research Information Manager (RIM). It was led by University College London[6]. It sought to define the need, explore the benefits and risks, and evaluate the approaches for establishing embedded research information specialists within multidisciplinary research environments. Researchers' requirements for support are mapped to existing services and infrastructure at UCL, in order to then define the RIM role skill set required to enhance the management of research information and data within the research group.

The project report sets out a methodology for others to adapt and use to define a RIM role to fit with their own specific needs through a combination of people, systems and services.

5 Themes identified as significant

5.1 Lack of integration

The lack of integration across institutional research information systems is increasingly recognised. It was strongly reinforced by the interviews carried out as part of the Imperial study: almost every institution cited a fragmented research system suite and a lack of integration as the fundamental problems facing them.

One of the original aims of the Imperial project was to examine ways of integrating reporting systems more effectively. Four institutions regarded their systems as 'well-integrated' (nine as partly integrated). However the report claims that, apart from Imperial College itself, none of the other institutions interviewed had mature enough systems to consider system integration. If this is indeed the case, more streamlined research information management is further down the line than anticipated. However it is possible that workarounds may found in the interim. For example during the BRIM project Huddersfield found that the EPrints publications repository and the HR database did not provide adequate web service interfaces for implementing the planned service-oriented architecture. This meant integrating systems at the data storage level instead (and not re-using processing capabilities); despite this compromise (including dependency on proprietary software from one supplier) it was still possible to demonstrate integration with and harvesting from existing institutional systems in a component-based way. 'A key result of the project was that it demonstrated how data stored in a University publications repository can be reconciled with other University computer systems to create accurate management information about staff publications.'

CRISPool was able to demonstrate that institutions with an existing integrated research system or CRIS have distinct advantages; for example, at a publications level, St Andrews was able to create CERIF-XML files directly from their PURE CRIS, whereas other institutional partners found this either extremely time consuming or impossible within the project timescale.

In addition to internal integration, several institutions, including some of the most research intensive involved in the Imperial study, cited the need to integrate internal application systems with external funder systems, such as the joint electronic submission system (JeS).

5.2 Research staff dissatisfaction

It was universally agreed by research-related staff taking part in the Imperial study that the current systems offering was unacceptable - administrative staff, and academics in particular had low satisfaction levels with the systems they used. This is directly linked to the system fragmentation issue discussed above. Most institutions interviewed were in the process of reviewing the systems used to manage one or more elements of the research cycle.

Usability continues to be a big issue. Interfaces are not necessarily intuitive. Academic staff underlined the need to design systems with easy-to-use interfaces and which were relevant to their needs. Access to the same systems is required at different levels of granularity

Most acknowledged that the RAE and the REF had become the primary drivers. As a result of this systems were reactive and focused upon corporate needs rather than meeting the needs of academic staff: 'our systems should be defined to run our business with the inevitable consequence that they will deliver what is needed for the RAE/REF'.

The UCL study also emphasised that while information and communication technology has a vital role to play in maximising the efficiency of administrative and research processes, *people* are required, to ensure that systems fit with researcher needs and workflows.

Participants in the Enquire project workshops stressed the need to minimise the burden on academics.

5.3 Capturing RIM requirements

The difficulty of articulating and agreeing upon RIM requirements surfaced many times in the Imperial report: RIM is very complex. (One result of this is seen to be a lack of investment in RIM systems – other more easily understood services take precedence.) The complexity is increased by the large number of institutional stakeholder groups involved with research information in its various forms.

It was often suggested that stakeholders wanted to specify perfect systems from the outset rather than focusing on core functionality first. This often led to poorly defined system projects which subsequently encountered difficulties with functionality, data cleansing and migration.

Institutions, (for a variety of reasons, some more valid than others), regard their research activities and research management activities as different from each other. This is not necessarily the case, but it creates problems for software suppliers.

As indicated, BRIM's main task was to capture RIM requirements at Huddersfield and create a prototype that integrated with existing business processes using standard interfaces. Interviews with RIM stakeholders were carried out in order to elicit requirements; given the six-month timescale it was decided to focus on the research project bidding process and the research publication data repository.

CRISPool found that taking time initially to define requirements (and prepare sample guidelines files) was very important, and this is generally the advice given. However this contrasts with the Imperial recommendation that institutions should not spend too much time on requirements since it allows less time to build the system.

A barrier is seen to be lack of shared standards in data and data definitions across the sector, which make it difficult to define systems requirements consistently across institutions. Imperial recommends a national framework for data standards which would enable institutions to specify generic systems.

5.4 Data quality

Data quality was identified as a big issue by the Imperial study interviews. High quality data is clearly the key building block for successful exchange of research information. However as an example, institutions report that the same data is often rekeyed into different systems; this increases the margin for error and decreases the likelihood of future interoperability (as well as being a waste of staff resources). Duplicated data in particular was recorded as a major source of dissatisfaction: 'Research office staff complained consistently of duplication of effort and poor data quality.'

Imperial recommends that resources for data cleansing, migration and conversion should be properly identified by a project and anticipated from the outset. This was one of the most consistent areas of difficulty in system implementation across the sector. The report also notes the absence of high level frameworks and standards in the UK with regard to data collection and data sharing.

At a practical level CRISPool highlighted the importance of CERIF sample files; these provided guidance for other institutions which were unfamiliar with CERIF, thus aiding accuracy and consistency. CRISPool also notes that the sample files from Glasgow and Edinburgh (person) required tidying up before they validated successfully against the CERIF-XML 2008-1.1 schema.

The topic of data was the single greatest cause for concern for researchers consulted as part of the UCL project. However in this context the issue was more about the difficulties of data management than quality, although the two can be closely linked.

5.5 Person identification

The issue of unique person identification was frequently cited across the projects, and is certainly not unique to RIM. The JISC Names project is working in this area.

At Huddersfield there is no centrally enforced uniform way of identifying members of staff across all university systems. One of the main problems occurs when members of staff enter co-contributors' details into the repository; many slightly different versions of the same name may be entered. As a consequence a lot of manual reconciliation of names is required before the information stored in the repository can be used to accurately provide statistics on the publication activities of staff throughout the university.

CRISPool partners agreed to create institution-specific unique identifiers for the organisations, persons and publications being brought together into CRISPool by using the UK Learner Provider number as a prefix to institutional identifiers. Enquire uses the Glasgow Unique Identifier (GUID) which is part of the author field of each EPrints record to link impact data.

5.6 CERIF

Of the five projects, only CRISPool and BRIM used CERIF; Enquire 'looked at' the standard, partly as a result of institutional involvement in CRISPool, but it was not part of the project plan.

At Huddersfield CERIF was used as a starting point in creating the 'RIMS' layer. Given the size of the standard, it was decided that the pilot implementation should focus on a portion of the CERIF model (information on staff research publications – using the EPrints repository). Fields were added to the CERIF model in order to accommodate local data. It was decided not to populate all fields, because some data was superfluous to local needs.

It is interesting that learning about CERIF carried an additional benefit, in allowing better understanding of local data structures:

The primary benefit we were reaping was that by adopting much of the structure of CERIF we were being guided in understanding and capturing the implicit structure of the data we are required to manipulate.

The CRISPool findings support the EXRI conclusion that CERIF should be used as the exchange format within the UK research information sector. The implementation of PURE has demonstrated the suitability of CERIF for capturing research information internally within St Andrews and Aberdeen. BRIM was also positive about the benefits of using CERIF, finding the overall structure of CERIF adequate for the functions of the layer that was prototyped. Huddersfield is also a pilot institution in the JISC CERIFy project, so is still taking forward CERIF implementation. Since St Andrews and Aberdeen have PURE installations, CERIF was already in use.

CRISPool did however find some problems using CERIF. The main technical issue was the fragmentation of CERIF-XML into many individual XML files; this means that processing is very resource intensive, since each item (whether person, organisation or publication) is defined by data in up to 10 related XML files. CERIF developers face a dilemma here, since the CERIF model needs to represent the real world of interrelated research information, but XML has a linear tree structure which cannot natively represent the complexity required. However, XML is also the vehicle of choice for data exchange in web services.

CRISPool was able to map most of the required data elements to the CERIF data model easily, with two exceptions:

- contact details – CERIF considers a person's contact details to be an attribute of the person, whereas in the CRISPool model they are an attribute of the relationship between

the person and organisation (since contact details change as a person moves jobs); this was worked around using classification

- URIs – CERIF supports a one-to-one relationship between publication entity and URI; CRISPool wished to record a DOI plus a URI for the full text version, but had to choose one only (DOI); this could also have been addressed using classification

CRISPool also worked closely with euroCRIS, and effected changes to CERIF 2008 based on UK requirements.

5.7 Research outputs and impact

Research impact and its recording is a key issue across UK HEIs currently. Work on impact and CERIF is continuing in the second JISC RIM phase.

The Enquire project in particular was affected by the Research Councils UK (RCUK) decision during 2010 to put the Research Outputs Project on hold. In recording information about impact for a range of research outputs, Enquire's focus had been on RCUK requirements. The project shifted focus from recording impact as part of the output to recording it as a separate entity associated with a staff member. It has not therefore been possible to deliver definitive output specifications that comply with RCUK requirements. However the project has instead produced generic specifications for impact and some key outputs which can be modified if/when different specifications are published.

The revised approach has included working with the JISC Institutional Repositories for Research Assessment (IRRA) project which had developed an EPrints RAE add-on, creating a separate mySQL database for recording measures of esteem, selecting publications and providing reports. Enquire has amended IRRA software, to focus on impact and esteem. The EPrints RAE add-on reporting function could be extended to an appropriate XML format such as CERIF.

5.8 Collaborative working

The Imperial report highlighted the lack of coordination between institutions as each implements their own solution to problems that are shared across the sector. It therefore recommended more collaboration (across both institutions and funders) in order to minimise duplication; also the establishment of network or body to facilitate institutional links and mapping of core processes.

The memory of the failure of the MAC initiative in the 1980s/90s (which attempted to get the sector to develop systems collaboratively) was often mentioned and several commented on the sector's inability to work together to address the underlying issues: 'Future efforts are tainted by past failures.'

However there are some examples of successful collaborations within the other RIM projects. For example useful cross-fertilisation was demonstrated by CRISPool which worked closely with JISC projects Enquire and ERIS; CRISPool also benefited from using R4R mapping advice and examples.

As discussed above, Enquire worked with the JISC IRRA project and further modified the EPrints RAE add-on developed by IRRA; the code is to be made available via the EPrints Bazaar app store.

6 Moving forward

Some of these themes are being taken forward by the JISC RIM2 projects, which have a more technical focus, based around increasing the takeup and use of CERIF in UK higher education. Further information and supporting documentation is available from the RIM pages at UKOLN[7].

7 References

[1] http://www.ukoln.ac.uk/rim/dissemination/2010/rim-cerif.pdf

[2] http://www.crispool.org/

[3] http://www.jisc.ac.uk/whatwedo/projects/rimsystems.aspx

[4] http://www.jisc.ac.uk/whatwedo/projects/enquire.aspx

[5] http://www.jisc.ac.uk/whatwedo/projects/rimtools.aspx

[6] http://www.jisc.ac.uk/whatwedo/projects/rimroles.aspx

[7] http://www.ukoln.ac.uk/rim/