Research Information Management in the UK: CERIF and metadata alignment

A supporting document for the JISC RIM Call, October 2010

Document details

Authors:	Rosemary Russell, UKOLN, University of Bath and Nikki Rogers, ILRT, University of Bristol
Date:	1 October 2010
Version:	1.1
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, the Museums, Libraries and Archives Council and the European Union. UKOLN also receives support from the University of Bath where it is based.

Research Information Management in the UK: CERIF and metadata alignment

A supporting document for the JISC RIM Call, October 2010

1 Introduction

1.1 Scope

The current document is designed to support the Research Information Management (RIM) strand of the JISC Call for proposals in October 2010. It therefore aims to provide an up to date picture of current RIM activities in the UK, with a focus on the CERIF standard and recent work towards alignment of metadata and vocabularies across research organisations.

Consequently the main body of the text focuses on current relevant projects and activities, together with the key organisations involved - including research funders and supporting agencies. The approach included telephone/Skype interviews with relevant stakeholders, as well as emailed queries to the RIM Group and information gathering from existing reports and documentation.

1.2 Background

Several existing RIM documents[1] provide a background to the UK research domain and the issues surrounding research information exchange. It is not proposed to summarise these here. However it is important to mention the context for the current document, and the expert group consulted.

The RIM Group of experts and stakeholders from universities, funders and representative organisations was convened by JISC in order to consider the results of the EXRI-UK (Exchanging Research Information in the UK)[2] report. EXRI recommended that CERIF should be the basis for the exchange of research information in the UK; it also recommended that work should be undertaken to align and map metadata semantics across the research domain. As a result several workshops and meetings were organised by UKOLN in July and August 2010 which examined the use of CERIF to map research information between institutions, RAE2008 and HESA; the Research Council's Research Outcomes Project (ROP) draft specification was also included, although there have since been significant changes to RCUK plans, as described in the funders section below.

2 Current projects and initiatives

2.1 Overview

As indicated, the focus of this document is tight, and this section describes only the particularly significant projects and initiatives in the context of this Call. Other documents and web sites describe related RIM projects. For example the UKOLN RIM page includes the five JISC RIM projects (with different foci) which were funded at the beginning of 2010, as well as other UK projects such as RMAS (Research Management and Administrative System)[3], funded by HEFCE

Rather than describe each key project in turn, project activities are grouped into three key areas, the first looking at data sharing across institutions using CERIF, the second REF-specific work and the third, project work which is institutional repository-focused. Relevant work being undertaken by research funders and other related organisations is also covered later.

The importance of CRIS working together with institutional repositories to exchange research information is recognised. Examples of other initiatives include the European Knowledge Exchange CRIS-OAR project[4] which has produced documentation to support interoperability between CRIS and Open Access Repositories; a CNR/euroCRIS workshop on CRIS, CERIF and institutional repositories was held in Rome in May 2010; and in the UK an event was held to explore synergies and opportunities for integration, also in May 2010.

2.2 Inter-institutional data sharing

Aggregating CERIF data from across multiple institutions is still very much at the pilot stage in the UK with two leading JISC-funded projects, R4R (Readiness4REF)[5] and CRISPool[6], offering case studies and demonstrators in this area. These case studies are producing valuable results in terms of successes, problems and issues involved in aggregating multiple institutional data using the CERIF data encoding standard.

2.2.1 CRISPool

The commercial supplier Atira recently wrote the aggregation system for CRISpool (the pooling of data for SUPA, the Scottish Universities Physics Alliance) with a resulting demonstrator - see the first version, search/browse demonstrator portal[7]. Some scalability issues were experienced (the performance of the XML aggregator database used by Atira declined rapidly as one university's - rather atypical in that it averaged 80+ authors per publication - data was integrated with the others). Visual design and usability testing work for this interface to the integrated data solution is still at an early stage. However, the demonstrator illustrates the advantage to Scottish universities of pooling their information in this way, and more institutions and pools are keen to join the initiative.

An informative final report has been delivered from the CRISpool project[8] (see References). To reach the point where the aggregator system could be developed, each participating university first had to undertake a conceptual mapping from its institutional repository/database schemas to the CERIF data model. Some CERIF 2008 classifications extensions and workarounds were necessarily devised as a result of this process. Next, CERIF-XML export files had to be created from each existing institutional database. Given the time and resource limitations of the project, a manual approach for this was used at some partner universities. This process, although straightforward from a technical point of view, was time-consuming, as described in the final report. It was quick and efficient in the case of St Andrews who have the PURE (commercial, CERIF-based) system fully deployed. The initial conclusions from the project on the suitability of CERIF as an exchange format are that it works well for mapping from the internal databases/schemas, because of the flexibility provided by the fragmentation of the model; however this very fragmentation causes scalability issues when aggregating data with many related entities e.g. publications with many authors. The project recommends continuing with the approach including work on solving these scalability issues in the interests of standardisation for information integration.

2.2.2 Readiness4REF

One of the goals of R4R is to demonstrate interoperability by automatically exchanging full (RAE-like) research data between institutions. At the time of writing in September 2010 the project has not yet achieved this, but developers are working on plugins for three repository platforms typically used for institutional research information (DSpace, EPrints and Fedora), with the project running until March 2011.

So far, Kings College London have exported from their CRIS to a REF and CERIF-based XML format they have developed - CERIF4REF (less complex than full CERIF). (The KCL CRIS can however also export to full CERIF format in addition to CERIF4REF.) Meanwhile the University of Southampton have been concentrating on extending their EPrints repository to be compliant with the CERIF data model. The Southampton work effectively offers a reproducible way for an institution to create a CERIF-CRIS from an EPrints repository. Southampton are currently planning to export CERIF4REF compliant data from this repository; the results could be two-fold in that they could firstly offer an example of how such a CERIF-CRIS can be helpful in generating REF data and secondly how this approach can be combined with KCL outputs, to demonstrate the interoperable sharing of institutional data. The project involves a similar effort regarding the creation of a CERIF-compliant DSpace plugin.

2.3 REF-specific

2.3.1 Readiness4REF

Although R4R concludes before final Research Excellence Framework REF requirements will be released, it has developed a CERIF4REF schema based largely on the earlier RAE 2008 format and with input from euroCRIS. This exercise has helped inform how CERIF may need to be extended for use within the contexts of both the REF and inter-institutional research information exchange.

R4R has undertaken case studies at six UK institutions (Goldsmiths College, University of London, Kingston University, University of Leicester, University of Reading, University of Ulster, University of the Arts) to produce in each case a thorough mapping of institutional data to CERIF2008. Keith Jeffery (euroCRIS) supported all six case studies, and there was also input from Anna Clements (CRISPool, St Andrews). The project's soon-to-be-released synthesis report will contain a useful summary of findings.

As mentioned above, R4R is also developing repository plugins at the Universities of Edinburgh, Southampton and Kings. These aim to demonstrate interoperability between institutions via use of the CERIF4REF common standard for exchanging data about research activity.

2.3.2 Symplectic software

Commercial vendor Symplectic[9] has been working to embed its information management software solution in approximately 17 institutions across the UK to date. Symplectic is involved with three JISC projects currently - ConnectedWorks[10] (a VRE project focusing on Sakai integration), the RePosit project[11] and the Dura project[12]. The Symplectic system is based on CERIF and its continued development focuses particularly on supporting the REF data collection format (as it emerges), and facilitating integration with institutional repositories (it has native compatibility with DSpace, EPrints, Fedora), thus potentially supporting the creation of REF submissions.

2.3.3 REF Impact pilot

In Autumn 2009, HEFCE conducted a REF Impact pilot[13] with several UK institutions. The results of this study are due in late 2010. Early investigations have shown that CERIF could be extended to cater partially for impact reporting in the REF, but there are no immediate plans to work on this. The research councils and other funders are also considering ways in which to collect and classify evidence of funded research impact.

2.4 Repository-focused

2.4.1 BRII: Building the Research Information Infrastructure

The BRII project[14] at the University of Oxford (now completed) focused on the use of semantic web technologies to share information about the institution's research activities. It has created a 'University Blue Pages' directory of researchers, interests and projects that captures information about research activity from a wide range of sources (including ORA, the institutional repository). The registry containing harvested data was constructed using the same architecture as the repository. CERIF was not used within the project, although the final report notes that 'the adaptation of the BRII registry to accept and distribute CERIF compliant data would be reasonably straightforward'.

2.4.2 Enquire: Enrich and Research Outputs and Impact

The University of Glasgow's Enquire project[15] has been using Enlighten, their institutional repository (which is EPrints-based), to record information about impact for a range of research outputs. Enquire initially used the draft requirements from the RCUK Research Outcomes Project to identify information to capture. Some extra fields have been added to allow the EPrints repository to record this data.

Initial work has been carried out looking at the RCUK/REF/HEBCIS/University of Glasgow entities and CERIF, and also exploring the export of impact data to the Medical Research Council's e-Val system.

2.4.3 ResearchRevealed

ResearchRevealed[16] is a two-year, JISC-funded project at the University of Bristol, due to complete by April 2011. Like the BRII project, it is piloting a semantic web solution to institutional research information integration, demonstrating how University central systems data about researchers, publications, organisational units and so on, can be linked to external online resources about the same things (for example, Web pages about collaborator researchers in other institutions or fuller details about grants from funder websites). ResearchRevealed's solution is particularly pertinent to the Web of Linked Data and a full demonstrator and open source software is being made available by the project. The project is also looking at the relationship of CERIF to a possible, harmonised, semantic web schema for the UK research domain - it has conducted a number of workshops on this theme, collaborating with the Universities of Oxford and Southampton. Finally, with the collection of 'impact evidence' in mind, the ResearchRevealed project is developing an Impact Bookmarklet tool[17].

2.4.4 OpenAIRE

The EU OpenAIRE[18] project has mapped the OpenAIRE data model to CERIF. The project is working on a solution to enable repositories and CRIS (including CERIF-based ones) to harvest and expose additional information to publication data via OAI-PMH – such as person, project, event and organisation entities related to publications. The work is based on experience with DIDL for compound objects in the Netherlands and the metadata format developed in the CRIS/OAR project. It is not yet completed.

2.5 Publishers

There are some UK examples of where institutions have developed automated solutions for importing purchased publications data to their institutional repositories - by contractual agreement with the publisher in question. For example, commercial supplier, Atira, wrote a plug-in for their PURE software to allow the system at St Andrews University to ingest publications data from Thomson Reuters (which provides an API that offers data in a proprietary XML format).

Thomson Reuters and Elsevier are members of euroCRIS and it is expected by euroCRIS that more publishers will join. Elsevier worked recently with Imperial College London on a JISC-funded RIM project[19]. At the project report[20] launch in August 2010, recommendations included addressing the lack of data consistency and (perceived) differences in research systems within the UK - this can include publication data as well as other, more metrics-based research information data. There are naming/identifier issues in terms of the integration of distributed publications data from suppliers - for example, although Thomson Reuters and Elsevier could be seen as usefully providing naming authority services for publications, they each have separate sets of publications identifiers.

3 Research funders

3.1 HEFCE (REF)

The REF system for data collection (and distribution of data to review panels) at HEFCE is currently under development. It is based on the previous system that was used for the RAE exercise, with current work focusing on core processes such as authentication mechanisms and workflow implementation, as opposed to Web forms for online submissions or data import/export formats. Meanwhile, the policy team continues to fix on REF data collection requirements and following finalisation of this, the specification for automated submissions by institutions will be decided upon and is likely to be published in 2011.

The R4R project has produced a 'CERIF4REF' XML file format as the potential basis of a core, standard, automated institutional REF return. REF Impact pilot projects findings will be available by the end of 2010 and will further inform the composition of the full REF submission. During 2011/2012 it should therefore be possible for institutions to test making their REF submissions either by Web form or via an XML-based, automated process. In the background to this, HEFCE have been exploring data alignment possibilities with RCUK and HESA, the key being to align definitions regarding the information being requested from institutions wherever possible.

3.2 Research Councils and CERIF-related activity

The Research Outputs Project (ROP)[21] had the goal of providing an integrated system for collecting and disseminating RCUK-funded research. However this system is now not to be implemented in the near term. A workshop was held in summer 2010 to investigate the suitability of CERIF 2008 for representing data collected by the Research Councils, in collaboration with euroCRIS. The general conclusion from the workshop was that in the vast majority of cases, the information that Research Councils collect from institutions maps well between Research Councils and could be modelled successfully using CERIF. Certain outputs (such as 'spin off company') are not explicitly modelled in CERIF, but euroCRIS were able to demonstrate how this can be easily accomplished as a role-based time-stamped relationship to the project, or other organisational units and/or persons.

Some further work was undertaken looking at mappings between the CERIF4REF XML format (produced by the R4R project) to the data schema developed in ROP. There was found to be a good match regarding descriptions of typical outputs (such as publications), with some non matches (for example regarding encapsulating career development).

Despite the fact that ROP is not proceeding as planned, there is likely still to be benefit in the Research Councils doing continued work in relation to the CERIF standard. However, there are no plans to hold another CERIF-related workshop at present.

A single system for research outputs information is to be adopted by four of the seven research councils, with a further two systems for the remaining research councils. One of these systems (MRC) is already a CERIF 2003 compliant system. Further work in relation to the harmonisation of data collection by the Research Councils could include trials with institutions with CERIF-compliant CRIS, submitting CERIF 2008-compliant reporting data as well as ingesting imports from RC systems, in order to test how easily this data could be exchanged by the systems in use.

3.3 The Wellcome Trust

Wellcome is currently considering the development of an online reporting tool to support the tracking of progress and accomplishments of those it funds; its approach is likely to have some similar functionality to the MRC e-Val system and is also likely to have some interoperability with UKPMC to harmonise the reporting burden on individual researchers. Wellcome is also in the process of developing a Data Warehouse to bring together data from across its internal systems; any online tool would link directly through to the Data Warehouse (and hence other data).

Currently there are no definite plans to implement CERIF in existing or future Wellcome data resources, although the facility to share data is recognised as desirable and a watching brief will be kept on CERIF-related development at HEFCE and other organisations

Wellcome is also involved in the Open Researcher & Contributor ID (ORCID)[22] initiative. This is a not-for-profit initiative that aims to solve the author/contributor name ambiguity problem in scholarly communications by creating a central registry of unique identifiers for individual researchers and an open and transparent linking mechanism between ORCID and other current author ID schemes. ORCID is a collaboration between research funders, publishers, data providers and researchers. The JISC Names project[23] is a participant in ORCID – it is developing a UK prototype name authority system for repositories.

4 Other related organisations

4.1 HESA: Higher Education Statistics Agency

During summer 2010 CERIF and HESA data experts met to evaluate the extent to which there are overlaps between the sort of information requested in a HESA return and the information that can be modelled using CERIF 2008. A detailed cross-mapping between the two data models was not attempted. The workshop concluded that although there are some clear overlaps on a few aspects, for example relating to the profiling of an institution’s funding sources, it does not look worthwhile to attempt to extend CERIF coverage to cater fully for HESA’s needs in the immediate (or even long) term.

Further work might usefully look at using a CERIF-compliant institutional CRIS to export data automatically to compile a subset of a HESA return where applicable, and to help inform how to ‘tighten up’ CERIF for that overlap with the HESA data model.

An institution’s data about programmes of study and teaching modules are usually stored or administered as part of systems and/or processes separate to the management of research information within institutions, hence there is not a large overlap between HESA reporting and reporting for the REF, say. However, where there is overlap (such as with expressing data about postgraduate students), there is an opportunity for harmonisation. HESA is planning discussions with HEFCE to look at alignment where HESA data could be re-used for the REF.

4.2 ARMA: support from UK research managers association

ARMA[24] is the professional association for research managers and administrators in the UK (90% of members are in universities). It provides professional development including training and information events (eg a recent event was a Repositories and CRIS workshop jointly hosted by ARMA and JISC), supports Special Interest Groups, and publishes Occasional Papers.

ARMA is supportive of the outcomes of the EXRI project and the JISC business case for CERIF in the UK:

> *Adopting CERIF as a UK wide standard for the exchange of research information will undoubtedly save valuable time and effort for research managers and administrators in supporting collaborative research proposals and projects. It will also enable better exchange of information with research funders and statutory bodies and opens up all sorts of opportunities for benchmarking. It also paves the way for research information to be properly integrated with the other core management information within research organisations.*

4.3 euroCRIS: providing support for CERIF

euroCRIS is the official custodian of CERIF. It supports and promotes CERIF and provides a growing range of documentation including the standard itself and tutorials (however there have been some requests for additional documentation, especially with regard to where UK requirements relate to the need for specific classification schemes). There is a biennial euroCRIS conference as well as biannual membership meetings. The core of the work is carried out in Task Groups. Membership fees are low and provide access to additional resources such as draft releases and newsletters as well as meetings and discussion fora.

There are now many euroCRIS members in the UK and they are generally willing to support CERIF-related initiatives by discussion or workshop attendance, in the interests of sharing knowledge, experience and good practice. CERIF-related activity in the UK has effected modifications to CERIF 2008 in the last year. Changes (including classifications) were requested by R4R and CRISPool on behalf of the UK community. They were approved by the CERIF Task Group and it is expected that the November 2010 release will include these changes as well as more detail on the funding entity. After that it is expected that there will be expansion of the facilities, equipment and services entities. UK members of euroCRIS are encouraged to attend member meetings (the next is in Prague in November 2010) to contribute views.

5 Directions

5.1 CERIF-CRIS growth in the UK

There is a growing momentum of CERIF-CRIS related activity in the UK, which the JISC Call in October 2010 will further encourage. Despite some frustrations experienced in the UK community over the complexity of the CERIF approach to using XML, the wider benefits and efficiency gains for the exchange of data are recognised. In parallel to the Call JISC is continuing to work with the RIM Group of experts and stakeholders in order to bring the many initiatives described together under a clear strategy.

As indicated, a number of UK players are already very actively involved in euroCRIS activities and recent euroCRIS newsletters have listed many new UK members. The growth in euroCRIS membership is unsurprisingly running in parallel to the installation of CERIF CRIS software at a number of institutions, eg in recent months Royal Holloway and University of York have installed PURE, and University of Hull, CONVERIS. Most recently, it was announced in September 2010 that the University of Stirling has also purchased CONVERIS. Several other institutions are in the process of tendering for systems.

Increased UK uptake means that more peer support for CERIF is likely to be available, as well as best practice guidelines for UK-specific requirements. The ability to use CERIF for REF submissions will be a further incentive.

5.2 CRIS beyond Europe

There are also indications that RIM/CRIS work is likely to become more international in scope – CASRAI[25] in Canada has proposed an international research metadata interoperability initiative, and hopes to involve US agencies, as well as euroCRIS, JISC and others. CERIF CRIS activity is already happening outside Europe, in the Middle East, South America and Canada (an institution in Mexico recently expressed interest in participating in a euroCRIS project).

6 CERIF contacts

The following people have indicated that they are happy to be contacted about their experiences of using CERIF:

- Anna Clements, University of St Andrews <akc@st-andrews.ac.uk>
- Mark Cox, King's College London <mark.cox@kcl.ac.uk>
- Keith Jeffery, STFC/President euroCRIS < keith.jeffery@stfc.ac.uk>

7 Acknowledgements

The authors are grateful to the many people who have contributed to the content of this document. The following people participated in one or more CERIF alignment workshops and meetings during summer 2010 and in addition many responded to detailed emailed queries. Those marked with an asterisk were also interviewed specifically for this document.

- Gordon Allan, University of Glasgow
- Les Carr, University of Southampton/EPrints
- Anna Clements*, University of St Andrews
- Mark Cox*, King's College London
- Michael Day, UKOLN, University of Bath
- Alan Green, STFC
- Bill Hubbard, CRC, University of Nottingham
- Neil Jacobs, JISC
- Neil Jefferies, University of Oxford
- Keith Jeffery*, STFC/President euroCRIS)
- Simon Kerridge, University of Sunderland/ARMA & RMAS
- Gerry Lawson, RCUK
- Scott Rutherford, HEFCE
- Andy Youell, HESA

Other interviewees

- Tim Brody, University of Southampton
- Richard Gartner, King's College London
- Gareth Edwards, HEFCE
- Daniel Hook, Symplectic
- Kevin Dolby, Wellcome Institute

8 References

[1] Links to key documents including JISC RIM information are available from the UKOLN RIM page: http://www.ukoln.ac.uk/rim/

[2] Exchanging Research Information in the UK: http://ie-repository.jisc.ac.uk/448/

[3] RMAS: http://as.exeter.ac.uk/rmas/

[4] KE CRIS-OAR project: https://infoshare.dtv.dk/twiki/bin/view/KeCrisOar/WebHome

[5] Readiness4REF: http://www.kcl.ac.uk/iss/cerch/projects/portfolio/r4r.html

[6] CRISPool: http://www.st-andrews.ac.uk/crispool/

[7] CRISPool portal: http://crispool.atira.dk/portal/

[8] CRISPool final report: http://www.st-andrews.ac.uk/crispool/media/crispool%20final%20report%20v2.1%20with%20appendices.pdf

[9] Symplectic: http://www.symplectic.co.uk/

[10] ConnectedWorks: http://www.caret.cam.ac.uk/page/jisc-collectedworks

[11] RePosit: http://jiscreposit.blogspot.com/

[12] DURA: http://jisc-dura.blogspot.com/

[13] REF Impact pilot exercise: http://www.hefce.ac.uk/research/ref/impact/

[14] BRII project: http://brii.bodleian.ox.ac.uk/

[15] Enquire project: http://www.jisc.ac.uk/whatwedo/projects/enquire.aspx

[16] ResearchRevealed: http://researchrevealed.ilrt.bris.ac.uk/

[17] Research impact tool prototype: http://researchrevealed.ilrt.bris.ac.uk/?p=69

[18] OpenAIRE: http://www.openaire.eu/

[19] Developing tools to inform the management of research and translating existing good practice: http://www3.imperial.ac.uk/research/jisc

[20] Developing tools final report: http://www.researchdatatools.com/downloads/2010-research-information-management.pdf

[21] ROP: http://www.rcuk.ac.uk/aboutrcuk/efficiency/Researchoutcomes/default.htm

[22] ORCID: http://www.orcid.org/

[23] Names project: http://names.mimas.ac.uk/

[24] ARMA: http://www.arma.ac.uk/

[25] CASRAI: http://casrai.org/