CERIF in Action: Synthesise, standardise and productionise CERIF for Higher Education Institutions

Anna Clements[a], Scott Brander[a], Valerie McCutcheon[b], Josh Brown[c], Dale Heenan[d], Thomas Vestdam[e]

[a] University of St Andrews, St Andrews, Fife, UK

[b] University of Glasgow, Glasgow, UK

[c] JISC Executive, London, UK

[d] Economic and Social Research Council (ESRC), Swindon, UK

[e] Atira A/S, Denmark

Summary

The CERIF in Action (CIA) project builds on previous work in the UK to explore the feasibility of adopting CERIF as a standard exchange format in order to improve operational efficiency and data quality across the research community. Unlike previous projects, CIA is looking at live operational systems and use cases relevant to all UK research organisations.

Whilst CERIF-XML itself is a standard schema, there is recognition that the UK research community is in danger of diluting the potential power of the standard and the benefits it can bring because of the varying ways in which CERIF has been mapped at the detail level in earlier projects. In order to be truly interoperable and maximise the efficiency savings that have been estimated if the UK were to adopt CERIF, agreement needs to be reached regarding the operational implementation of CERIF, e.g. which entities and semantics to use.

1 Introduction – the CERIF in Action project

1.1 Background to the project

CERIF has been recommended for adoption across the UK research community in order to streamline the flow of data between stakeholders, improve data quality and reduce costs across the sector by enabling more efficient and effective procedures related to research information data exchange.

The project is part of the large investment made in improving research information management in the UK by the Joint Information Systems Committee (JISC)[1], the UK's expert on information and digital technologies for education and research. JISC is funded by all 4 UK higher education funding councils and their Research Information Management[1] programme of work, encouraged and supported by euroCRIS, has been instrumental in catapulting the UK to the forefront of national CERIF adoption programmes.

[1] For more information, see www.jisc.ac.uk

Several successful JISC-funded projects have spearheaded this adoption process and through working closely with euroCRIS, have led directly to enhancements to the CERIF standard for the UK community. Examples include a recommendation from the CRISPool[2] project to reduce the fragmentation of the CERIF-XML standard and the extension of the CERIF model to include Impact measures from the MICE[3] project.

However, although CERIF is a standard model, there is recognition that the community is in danger of diluting the potential power of the standard and the benefits it can bring because of the varying ways in which CERIF has been mapped at the detail level in these separate projects.

In order to be truly interoperable and maximise the efficiency savings that have been estimated if the UK adopts CERIF, we need to agree precisely which CERIF entities we use, how we code identifiers and the semantics for the various entities and their roles and relationships to one another. We also need to demonstrate whether CERIF-XML works in key exchange scenarios and so begins to realise the cost-savings and improved data quality forecast. The CERIF in Action project aims to address these issues.

1.2 Timescale and partners

CERIF in Action (CIA) is a 12 month project which started in November 2011. It is led by the University of St Andrews and includes research organisations, funders, system suppliers and CERIF experts. The university partners joining the University of St Andrews are the Universities of York, Glasgow & Edinburgh and Trinity College Dublin. Research Councils UK[4], the strategic partnership of the UK's seven Research Councils who invest around £3 billion in research per year, represents the funder perspective. The three leading suppliers of current research information systems (CRIS) and institutional repository (IR) software in the UK are involved as commercial partners: Atira[5], Symplectic[6] and ePrints[7]. Finally, euroCRIS provide expert advice and support to the project.

1.3 Key objectives

The key objective is to extend existing live systems at research organisations and funders to use CERIF-XML to exchange research information and thereby improve the efficiency and effectiveness of these organisations by removing data entry duplication. CIA meets this objective by addressing the following use cases:

- **Use Case 1: Researcher moving to another research organisation**: The first use case addresses the current situation where a researcher has to re-enter research activity information when moving to a new Institution leading both to duplication of effort and increased opportunity for data errors.

[2] For more information, see www.crispool.org

[3] For more information, see http://mice.cerch.kcl.ac.uk

[4] For more information, see http://www.rcuk.ac.uk/Pages/Home.aspx

[5] For more information, see http://www.atira.dk

[6] For more information, see http://www.symplectic.co.uk/

[7] For more information, see http://www.eprints.org/services/

The supplier partners build production-ready plug-ins using the agreed standardised CERIF-XML for their products. These plug-ins are used by institutional partners to exchange data between their live CRIS or IR systems and thus reuse the same information

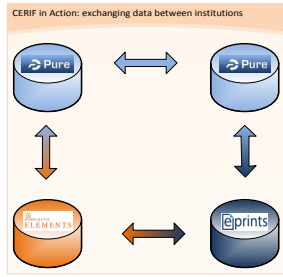

Figure 1: Exchanging data between institutions

- **Use Case 2: Researcher or research organisation uploading data on research outcomes to research funder at end of project:** This use case addresses the new requirement for researchers to provide evidence of the outcomes from a research project to the Research Council funder.

  The same plug-ins as above will be used to export data from the research organisation's (RO) CRIS or IR and then import into the RCUK ROS system. This will lead to the RO remaining the authoritative source of this information which it can then reuse for internal management and extend to other non-RC funders.

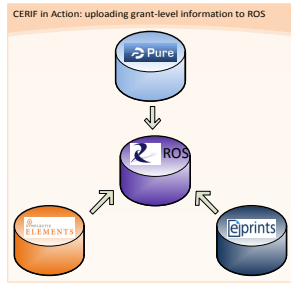

Figure 2: Uploading grant-level information to ROS

1.4 Key deliverables

- A standard CERIF-XML model suitable for exchange of information on people, organisations, publications (text-based), projects and related funding. The 12 month project length and the requirement to produce production-ready project deliverables, as opposed to proof of concepts or pilots, means we have limited the scope of CIA to these particular CERIF entities.
- Plug-ins for Pure, Symplectic and ePrints to enable export and import of the research information data defined above. All of these suppliers have agreed to embed these plug-

ins into their standard product at no additional cost to existing customers and to continue to maintain the plug-ins as part of their standard business model. The project aims to demonstrate the use of these plug-ins in live production environments.

- A roadmap for adoption of this standard CERIF-XML - including an evidence-based report on the recommended steps and the expected costs and benefits of embedding and extending the CERIF in Action approach to other Institutions and to other key data exchange scenarios.

2 The national UK perspective

Higher Education research in the UK has a major role that reaches a long way beyond our universities. Government research spending (£4.6 billion [1], or approximately €5.5 billion, of government funding in 2012 for instance) is complemented by funding from charities (such as the Wellcome Trust) and private companies to produce a large and vibrant research community. The UK's share of research papers published and citations (6.4% and 10.9% respectively [2]) demonstrate the intellectual impact of this community. However, the management of information about that research base has traditionally been fragmented and information has been hard to gather and analyse. It is clear that any efficiency gain or improved management in an area of activity of this size has the potential to free up both money and staff time that could be put back into the core activity of research itself. The focus, therefore, has been on linking 'silos' of research information and improving the efficiency of research information exchange. JISC commissioned the "Exchanging research Information in the UK" (EXRI-UK) report in 2009, which identified CERIF as the best option for improving the interoperability and exchange of research information ([3] Rogers et al.).

This led to work to quantify potential savings from the use of CERIF ([4] Bolton, 2010), and a cycle of investment in CERIF projects[5] (now totalling more than £2 million) which have advanced the use of CERIF in the UK and created a pool of expertise and practitioners who have materially contributed to the evolution of the standard. This has led to increasing engagement with commercial software vendors and the rapid growth in the procurement of CERIF research management systems (from zero to 30% of UK universities within 3 years) with funders and other central bodies increasingly engaging with and implementing the standard [6].

At the national level, concerted effort to introduce CERIF to the UK has gone well, and continues to gain pace, although not without problems and challenges. The Higher Education sector is extremely diverse, and consistency is vital if the power of CERIF is to be harnessed to support our national research base. The CERIF in Action project will help with maintaining consistency of implementation and in extending the use of CERIF-XML into operational systems.

Membership of euroCRIS, the European organisation responsible for the maintenance of the CERIF standard, has a large UK contingent, with 24 organisations as demonstrated in Figure 3 below:

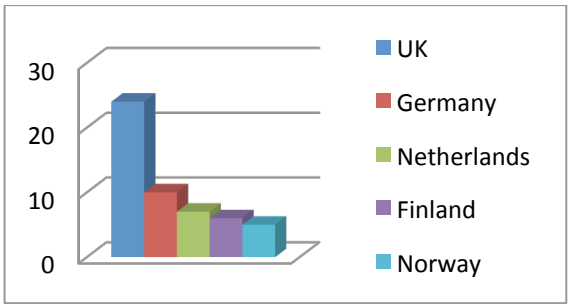

Figure 3: euroCRIS institutional membership by country (reproduced by kind permission of R. Russell, UKOLN)

3 The funder perspective

3.1 Introduction

As the public bodies charged with investing tax payers' money in science and research, the Research Councils take very seriously their responsibilities in making the outputs from this research publicly available – not just to other researchers, but also to potential users in business, Government and the public sector, and also to the public.

The Research Councils are committed to the guiding principles that publicly funded research must be made available to the public and remain accessible for future generations and, following a wide consultation with stakeholders, RCUK published a position statement on access to research outputs[8].

3.2 Research outcomes system

The Research Outcomes System (ROS) is a web-based system, launched in November 2011, which allows users to provide research outcomes to four of the eight Research Councils: Arts & Humanities Research Council (AHRC), Biotechnology and Biological Sciences Research Council (BBSRC), Economic and Social Research Council (ESRC), and Engineering and Physical Sciences Research Council (EPSRC). Additionally, a fifth Research Council, NERC, and will be collecting research outputs data via ROS from 2013.

The ROS can be used by these Research Council grant holders to input outcomes information about their research. It can also be used by Higher Education Institution (HEI) research offices to input research outcomes information on behalf of grant holders and/or access the outcomes information of grant holders in their institution.

CIA will contribute to this effort greatly as it is anticipated that it will enable the HEI as a whole, or the individual researcher, to export this information automatically rather than re-enter it into

[8] For more information, see http://www.rcuk.ac.uk/research/Pages/outputs.aspx

ROS, saving time, cost, opportunity for error and duplication of effort; thus ultimately allowing the researcher to do their core task: research.

The information gathered by ROS is key to the Research Councils strengthening their evidence base for strategy development, and crucial in demonstrating the benefits of Research Council funded research to society and the economy.

4 The research organisation perspective

4.1 What CERIF in Action addresses

The current method for Research Organisations to provide data to the RCUK Research Outcomes System (ROS) is via bulk uploading spreadsheets. Obtaining information from core systems and mapping the fields to those provided and defined by RCUK is extremely time consuming, error prone, and open to interpretation. Research Organisations are concerned to provide quality information to the RCUK in a timely and efficient manner. We hope that utilising the CERIF-XML schema and semantics defined by the CIA project to deliver tight data specification and set up automated data export and import routines will help save public money as the administrative costs of delivering the information will be reduced. Having clearer definitions will also reduce the potential for confusion and anxiety within the academic community and for administrators interacting with the ROS system.

The same logic applies to exchanging data between Research Organisations; for example, when a researcher moves organisation and wishes to take details of research activity such as awards and outputs with them to populate web profiles at their new employer.

4.2 What CERIF in Action could address in the future

We see that once the use of a CERIF-XML standard is embedded into our core research information system/s, we have the potential to extend the model to new scenarios and thereby reduce effort across the sector by reusing a standard format. The major concern at the moment is the lack of basic authority data sources to facilitate efficient data exchange. An example is the lack of an agreed standard list of UK research organisations and research funders. Work needs to be done to allocate responsibility for the maintenance of such data sources with accompanying machine readable services.

Some areas into which the CIA standard could be extended:

- Statistical returns e.g. HE & business/community interaction survey (HEBCIS)[9]
- Pooling/aggregation of research activity across organisations; in Scotland this is an important use case and it is becoming increasingly important across the rest of the UK as more and more institutions seek to collaborate and share resources within geographical or disciplinary areas.

[9] For more information, see http://www.hefce.ac.uk/econsoc/buscom/hebci/

5 Software supplier perspective

5.1 What does CERIF compliance mean?

Software suppliers, or CRIS vendors, are often faced with CERIF compliance requirements, however it is never qualified what that actually means. If "compliance" is taken in the context of the ability to exchange information stored in a CRIS via CERIF (-XML) with another CRIS, then this requirement would still be quite vague.

Whilst CERIF-XML is a standard, a real "standard" for how to actually exchange information using CERIF-XML is still missing. By using the CERIF standard, you can solve most mapping issues from the standard to a proprietary CRIS format (i.e. the internal data model of a given CRIS). Furthermore, the CERIF standard does provide a set of "standard" classifications.

However, mapping to and from CERIF can be done in many ways - there is a lot of "modelling" freedom, and a more formal specification (or patterns) on how to model is lacking. Secondly, the set of "standard" classifications has never been put to the test, and as a consequence it is not yet known whether the set of classifications covers the information needs of real scenarios. Hence, certain business rules or the specific model of one CRIS may result in a CERIF output that is semantically different from the output of another CRIS, thereby making exchange impossible. The only way to alleviate such problems it to define a standard for exchange that covers all aspects of exchange: mapping, semantics, and protocol. Furthermore, the standard must be rigorous, comprehensible, exhaustive (no interpretation need), as well as tested and proven.

There is therefore a need to experiment with CERIF, or specifically CERIF-XML, as an exchange standard in order to arrive at a fully comprehensible standard.

5.2 How CERIF in Action can help

An important aspect of the CIA project is that it involves three vendors and ROS, the RCUK research outcomes system. This will allow for challenging exchange scenarios that will only succeed if the different parties are actually using the same standard and are interpreting the semantics defined in the standard in the same way.

Obviously the usefulness of such an exchange format goes far beyond unqualified customer requirements to CERIF compliance and CRIS-to-CRIS exchange. CRIS vendors can support many interfaces to systems external to their CRISs; for example, "publication databases" (Scopus/SciVal, Web of Knowledge/Incites, PubMed, etc.), public repositories (e.g. the FRIS Portal in Flanders, a local DSpace, EPrints repository) and official evaluation systems (e.g. REF in the UK and FI in Denmark). If all such systems were using the same exchange format (at least when it comes to the raw payload data), then maintenance of such external interfaces could be easier, and building interfaces for new external services could be faster.

6 Project Progress

6.1 Project progress and interim outcomes

The project started with a review of previous projects where CERIF-XML models have been produced, identifying similarities and differences in approach, issues and how they were solved, standard (and non-standard) vocabularies used, and the existence (or lack) of relevant data authority sources. This was completed with a detailed mapping[10] between projects which highlighted a number of areas where the projects differed. Other issues that arose were workarounds which had been done to complete the relevant projects and semantic issues.

A further mapping was done between CERIF and the Research Outcomes System (ROS) system to highlight any areas where there may be issues transferring the data between a CRIS/IR and ROS.

These issues were discussed with the project partners and the wider CERIF community, and a number of bespoke classification schemes were devised to sufficiently enable the exchange of information, including a combination of new classification schemes and extensions to existing CERIF classification schemes.

A functional specification[11] has been received by the software partners for development of the plug-ins. A sample CERIF-XML file has also been produced which validates against the CERIF 1.4 schema[12] and follows the embedded structure newly adopted by euroCRIS. This sample XML file also helped identify further areas that may require resolving in CERIF and a report will be produced for recommendation to the CERIF Task Group.

6.2 Lessons learned and implementation issues

Much of the attributes of the entities mapped across to the CERIF model easily and, where possible, formal CERIF semantics were used. However, there were a number of instances where executive decisions and workarounds had to be made:

6.2.1 Persistent universal unique identifiers (UUIDs)

One key challenge in this project, as with others, is unambiguously identifying people, universities & departments, funders, projects, and so on.

A number of possible solutions to this issue were discussed in the early stages of the project and work being done as part of the JISC NAMES[13] and ORCID[14] projects will hopefully solve the person name ambiguity problem at least by creating a registry of unique identifiers for individual researchers. However, neither project is as a stage where it is implementable in CERIF in Action.

[10] Mapping spreadsheet available at http://sdrv.ms/HrvV4b

[11] Functional specification available at http://sdrv.ms/HdKg1Q

[12] CERIF 1.4 XML Schema: http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.4/CERIF_1.4_0.xsd

[13] For more information, see http://www.jisc.ac.uk/whatwedo/programmes/inf11/shainf/names2.aspx

[14] For more information on ORCID, see http://about.orcid.org/

The CIA project has therefore used randomly-generated Universally Unique Identifiers (UUIDs) to identify the majority of items, with the exception of Persons and Organisations which will use a concatenation of the institution and an internal identifier for internal items; and a UUID for external items.

Duplication handling will therefore need to be performed by each of the import systems to identify duplicates and allow de-duplication or merging, and/or author matching as appropriate. The UUIDs will be generated "on the fly" when first extracted from the source CRIS/IR but are retained for reuse thereafter. The use of UUIDs may change in future if a definitive source of UUIDs for any particular entity becomes available. This process will rely on the de-deduplication capabilities of the system importing the data and, in general, the more data that is available, the more successful the de-duplication process will be.

A future release of CERIF (provisionally 2.0) may include federated identifiers, with the ability to store, alternate identifiers with a certain entity. For the purposes of this project, however, new classification schemes were created to enable the recording of alternate identifiers, for example, Web of Science ID, Scopus ID, and so on.

6.2.2 Separation of local and standard semantics

There are instances where it has not been possible to map to standard CERIF semantics directly and this has led to a tension between losing the richness of the data by mapping to the standard CERIF semantics or compromising the interoperability between the systems. Examples of this include personal job titles and organisation units. For the purposes of this project, we have chosen the former but recognise the need for further work in this area.

6.2.3 Classifying non-CERIF semantics

A number of bespoke classification schemes had to be created for CIA to allow true interoperability and some or all of these will be recommended to the CERIF Task Group for inclusion in future versions of CERIF. The bespoke classification schemes are either extensions to existing CERIF ones (such as a list of publication types) or new ones required for exchanging the data effectively (for example, personal [honorific] titles).

7 Conclusions

The CIA project has successfully built on previous CERIF work in the UK and is expected to result in a considerable expansion in the use of CERIF-XML by virtue of the involvement of the main software providers and the key funder stakeholder – RCUK. This is expected to result in efficiency improvements across the sector and considerable improvement in data quality and availability. However it has also clearly demonstrated that for the full benefit of the standard to be realised, the community still has more work to do to address the two main issues described in 6.2 :

- Unambiguous identification of entities
- Common language/vocabulary/semantics – maintained by recognised authority sources

References

[1] http://www.jisc.ac.uk/whatwedo/themes/informationenvironment/researchinfomgt.aspx [Accessed 14/02/2012]
[2] http://www.bis.gov.uk/assets/biscore/science/docs/a/10-1356-allocation-of-science-and-research-funding-2011-2015.pdf
[3] http://www.bis.gov.uk/assets/biscore/science/docs/i/11-p123es-international-comparative-performance-uk-research-base-2011-summary.pdf
[4] Rogers, N. and Huxley, L. and Ferguson, N. (2010) *Exchanging Research Information in the UK: final report.* Project Report. Available at: http://ie-repository.jisc.ac.uk/448/1/exri_final_v2.pdf [Accessed 13/2/12]
[5] Bolton, S. (2010) *Business case for the adoption of a UK standard for research information inter-change*. Report to JISC. Available at: http://www.jisc.ac.uk/media/documents/publications/reports/2010/Businesscasefinalreport.pdf [Accessed 13/2/12]
[6] http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement.aspx
[7] http://www.ukoln.ac.uk/isc/reports/cerif-landscape-study-2012/CERIF-UK-landscape-report-v1.1.pdf
[8] Russell, R. (2012) *Adoption of CERIF in Higher Education Institutions in the UK: A Landscape Study* http://www.ukoln.ac.uk/isc/reports/cerif-landscape-study-2012/CERIF-UK-landscape-report-v1.0.pdf [Accessed 17/03/12]

Contact Information

Anna Clements
Enterprise Architect
IT Services. University of St Andrews
Butts Wynd, St Andrews, UK
KY16 9AL
akc@st-andrews.ac.uk

Scott Brander
Research Data Project Manager
IT Services. University of St Andrews
Butts Wynd, St Andrews, UK,
KY16 9AL
scott.brander@st-andrews.ac.uk

Dale Heenan
Web Project Manager
Economic and Social Research Council
Polaris House,North Star Avenue
Swindon, UK
SN2 1UJ
Dale.Heenan@esrc.ac.uk

Valerie McCutcheon
Operations Manager,
Research and Enterprise
University of Glasgow
10 The Square,
Glasgow, UK, G12 8QQ
Valerie.McCutcheon@glasgow.ac.uk

Josh Brown
Programme Manager, Digital Infrastructure
JISC Executive
Brettenham House
5 Lancaster Place
London, UK
WC2E 7EN
j.brown@jisc.ac.uk

Thomas Vestdam
Project Manager
Atira A/S
Niels Jernes Vej 10
9220 Aalborg Oest
Denmark
tv@atira.dk