CERIF-CRIS FOR THE EUROPEAN E-INFRASTRUCTURE

***K Jeffery*[1*] *and A Asserson*[2]**

[*1]*Science and Technology Research Council, Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Chilton, Didcot, Oxfordshire OX11 0QX, UK*
Email: keith.jeffery@stfc.ac.uk
[2] *Department of Research Management, University of Bergen, Nygardsgaten 5, Bergen, Norway*
Email: anne.asserson@fa.uib.no

ABSTRACT

The European e-infrastructure is the ICT support for research although the infrastructure will be extended for commercial/business use. It supports the research process across funding agencies to research institutions to innovation. It supports experimental facilities, modelling and simulation, communication between researchers, and workflow of research processes and research management. We propose the core should be CERIF: an EU recommendation to member states for exchanging research information and for homogeneous access to heterogeneous information. CERIF can also integrate associated systems (such as finance, human resource, project management, and library services) and provides interoperation among research institutions, research funders, and innovators.

Keywords: Information communication technology, Standards, Data model, Metadata, Formal syntax, Declared semantics

1 INTRODUCTION

The e-Science concept, developed in the UK from the paper by Jeffery (1999a), is now internationally widespread as e-Research. The concept is based on the idea that the end-user, wherever located in the research environment, should have homogeneous and easy-to-use access to all systems in the research environment covering: a) experimental set-up and control; b) data taking and analysis; c) appropriate modelling and simulation; d) comparison with experimental results; e) cooperative working and researcher interaction; f) scholarly publication; g) curation of research datasets and software; h) innovation; i) via the media providing information to the general public, all related to a framework of research management by funding agencies and research institutions including research proposals, project management and reporting, and financial and human resources management. This concept is closely related to the utilisation of research facilities and equipment (detectors, instrumentation) where European strategies are being developed, e.g. by ESFRI. Such a vision demands an infrastructure (the e-infrastructure) to support the whole range of research activities catalogued above, from end-to-end of the research process and across all research disciplines. The requirement is for an integrated e-infrastructure to support the end-user to minimize effort and to maximise effectiveness and efficiency.

At present, the ICT systems for experimentation and data collection and analysis, for simulation and modeling, for intercommunication among researchers, and for management of research in both funding agencies and research institutions are heterogeneous. There is clearly a pressing need for them to interoperate to improve effectiveness and efficiency. Furthermore, there is a need for the information to be open access (freely available at the point of use, see the Wikipedia definition: http://en.wikipedia.org/wiki/Open_access) to increase productive use of the information and to speed the research and innovation cycles. This openness requires management, however, such that publication precedence, IP (intellectual property) management, and certain security restrictions are maintained (Jeffery & Asserson, 2005).

In order to achieve the open interoperation, but with managed restrictions, two things are required:
1. metadata to describe all the systems, services, persons, organisations, projects, etc. involved and
2. a standard method of intercommunication among the components with a formal syntax and declared semantics (to permit machine-understandability as well as machine readability).

For this, we propose CERIF (Common European Research Information Format), a EU (European Union) recommendation to member states that forms a definition for the data structure of a CRIS (Current Research Information System) or for the data structure to be exported/imported by an interoperating CRIS.

The rest of the paper is organized as follows: Section 2 defines a CRIS (Current Research Information System) and the requirements of users in the research environment, Section 3 describes the emergence of CERIF, Section 4 outlines its use for integrating systems within an organisation and between organizations, and Section 5 concludes.

2 CRIS AND REQUIREMENTS

Research funding organisations require CRIS to manage funding programmes, applications against a funding programme, awarded grants (projects), and the results (products, patents, publications) associated with those funded projects. In addition, they need to keep track of persons (in various roles such as researchers, reviewers, research managers) and organisational units (such as research institutions, faculties, departments, groups, units, commercial companies, government departments).

Research institutions need to keep track of funded projects and associated persons, organisational units, and the outputs produced as well as facilities and equipment used, services provided, and events held.

Researchers need to propose research projects, set up experiments, research the literature, intercommunicate, analyse data, perform simulation and modelling, produce publications, and where appropriate, produce products and patents, transfer knowledge for innovation, and interact with management.

Innovators need to understand the outputs produced from research but also the context of that research and any IP issues.

The media (and hence the general public) need good ‘research stories’ relating the research to its effect on everyday life; hence the media need access not only to the outputs but also to the contextual data about the project, persons, organisations, etc.

Several things are noticeable immediately:

1. there is a large intersection of required information among the major kinds of organisations and persons involved in research and
2. the inter-relationships among persons, among organisational units, among projects and other research entities are not hierarchic but many to many, i.e., the relationships require representation by a fully connected graph.

Historically all the major kinds of research information users have developed their own ‘home brew’ systems. This is particularly true of systems managing experimentation and simulation / modeling and those managing research and managing researcher intercommunication including scholarly publications. More recently, some have started to use commercial ERP (Enterprise Resource Planning) software for research management and adapted it to research use. Similarly, there are a limited number of systems for managing scholarly publications. Nonetheless, there is great heterogeneity, which hinders effective and efficient management of research and research communication. There is a need for a standardised CRIS as argued in Zimmerman & Jeffery (2004) or, minimally, a standardised intercommunication among research organisations.

3 CERIF

It was against this background that ideas concerning interoperation of systems supporting research have progressively developed. Particular research communities have developed de facto standards for interchange of experimental data, an early set of examples date back to the International Geophysical Year 1958. Since the early 1990s, the idea of intercommunicating scholarly publications electronically has developed with the open access movement. The idea of intercommunicating research management information dates from the early 1980s as catalogued in Asserson, Jeffery, & Lopatenko (2002) with various initiatives culminating in the EC (European Commission) drawing together a team consisting of one government-nominated expert per member (and associated)

state to define a format. That format is CERIF; since 2002 the EC has handed the development and maintenance of this standard to euroCRIS (www.eurocris.org), a not-for-profit organisation.

CERIF is a data model; euroCRIS makes its definition available through open access, but additionally, euroCRIS members have access to scripts to set up a CERIF-CRIS and to the expertise of the CERIF Task Group members within euroCRIS. CERIF can be used to build a new CRIS or to wrap an existing CRIS to allow interoperation with other CERIF-CRISs or CERIF-wrapped CRISs.

The data model consists of entities (objects) of interest to the domain of research and relationships between them. The major entities are projects, persons, organisational units, and results (which may be products, patents, or publications). There are entities for funding, facilities equipment, services, events and skills, & CV, each with defined attributes. CERIF thus has a formal syntax for its data structure.

Within entities, text-based attribute instances can be multilingual (and flagged as automatic or human translated as well as recording the language), and the use of Unicode means that CERIF can represent all written characters. In addition, CERIF provides facilities to store authority files of approved terms for classification, e.g., the UNESCO Spines thesaurus or the UDC (Universal Decimal Classification) and closed lists of valid values for certain attributes (e.g., the ISO country codes). These classification terms and closed list terms are stored within CERIF in such a way that there may be multiple (multilingual) classifications or closed lists for any attribute; this provides the declared semantics of CERIF.

CERIF thus satisfies the first requirement of a data model: to represent the domain of research activity with formal syntax and declared semantics.

Relationships can be between instances of any two entities or between two instances of the same entity. An example of the first is the relationship between a person and a publication. All relationship instances are characterised as being between two date/time points (which may be the same indicating a single point in time or null indicating unknown) and a role, for example, ‘author’. At the same time interval, there may be a relationship instance between the publication and an organisational unit, for example with the role ‘publisher.’

An example of the second is the relationship between one organisational unit and another, for example, with the role ‘is part of’ (representing, e.g., a department of a university). It should be noted that the relationships are n: m (many to many) thus there can be many persons as authors of one publication and many publications with the same person(s) as author(s). This flexibility also allows an organisational unit of, for example, the type ‘research centre’ to be ‘owned’ by two or more departments within one university.

CERIF thus fulfils the second requirement: being able to map a fully connected graph of relationships and moreover to provide time stamped relationships thus allowing the state of the domain of interest at an earlier time to be reconstructed.

Documentation on the model including specifications and tutorials can be found at www.eurocris.org/cerif/introduction. There are now many implementations of CERIF, variously following successive CERIF versions but all capable of interoperation of the core entities and relationships. Indeed, there are commercial ICT companies offering CERIF solutions.

4 CERIF FOR INTEGRATION

CERIF has the capability to integrate systems within one organisation and also to permit interoperation among organisations, either by data exchange or by providing homogenous access over the heterogeneous systems of other organisations.

4.1 CERIF in one organisation

A typical organisation engaged in research, whether a funding organisation, a research institution, a publisher, a commercial company wishing to use results for innovation, or a media company, is likely to have various ICT

systems already in operation, e.g., for finance, human resource management, or project management. It is likely that they do not interoperate very well, that research domain information is stored multiply with corresponding problems of integrity maintenance, and that the representation of the research domain is inconsistent across the various systems in syntax and semantics.

CERIF can be used as an integration among these legacy systems to provide a uniform view of the research domain that can then in turn be used to represent the research organisation to others. The information on finance in CERIF can be used to access detailed information in a finance/accounting system; similarly the information on a person in CERIF can be used to access more detailed records in the human resources system. Analogously, the same technique can be used for projects, facilities, equipment, events, and so on. CERIF is thus being used as homogeneous wrapping metadata, as discussed in Jeffery (2000), to achieve this integration.

An important aspect is that CERIF information provides access to more detailed information on products, patents, and the full text or multimedia of publications. The requirements concerning products are solved by using CERIF as metadata to describe, and provide discovery of and access to, the products of research, such as datasets and software. This is a core area of e-Research, and CERIF allows not only appropriate access but also provides the research contextual information so that the e-Research work can be understood more readily and the end-user supported appropriately. Similarly, CERIF provides access via metadata to detailed patents records. The requirements of, and solution for, publications have been discussed extensively in Asserson & Jeffery (2004; 2005), following initial work by Jeffery (1999b), while the overview of the linkage among a CERIF-CRIS and repositories of research datasets and software and of publications was considered in Jeffery, Lopatenko & Asserson (2002), Jeffery (2005b), and Jeffery (2007).

CERIF can do more as described in Jeffery & Asserson (2006a) and developed further in Jeffery & Asserson (2008): it can be used as the primary source of organisational relationships, persons, and projects, such that it can populate user provisioning (security and authorisation information, e.g., for workflow), directories (such as Microsoft Active Directory or LDAP), and web pages describing the organisation, its structure, its work (projects), and its employees.

Finally, CERIF can support all the steps (workflow) of the research process within an organisation as explained in Jeffery & Asserson (2006b), thus supporting e-Research while maintaining the full information on the context of the research work.

4.2 CERIF between organisations

CERIF can provide the link to allow interchange of data and information from one organisation to another. By providing a canonical format, with formal syntax and declared semantics, CERIF converts the [n*(n-1)] intercommunication problem into a [n+1] problem. Each organisation wishing to intercommunicate is required to export information in CERIF and to be able to import, internally converting from or to the local legacy systems.

A slightly more sophisticated technique allows a query to access multiple organisations. The query is framed using CERIF entities and attributes, at each organisation, the query is translated into the query language (with syntax and semantics) of local legacy systems, and the results of the query are translated to CERIF for export. Thus the end-users think that they are accessing a CERIF system (with query and homogeneous response); yet in fact, they are accessing multiple heterogeneous systems.

Both techniques have been proved in demonstrator projects, and a description of the architectural elements is published in a technical report by Jeffery (2005a).

5 CONCLUSION

It is clear that CERIF can represent the domain of research information, e-Research, with a formal syntax and declared semantics. This permits integration of ICT systems in one organisation and interoperation across multiple organisations. CERIF can thus act as the core of the e-infrastructure, supporting all aspects of research activity.

Furthermore, this architectural concept can be utilised generally in business, environmental protection, healthcare, e-learning, culture and leisure, and many other application domains.

6 ACKNOWLEDGEMENTS

This paper draws heavily on the work of the euroCRIS community, and their contributions are here acknowledged. Special mention should be made of Brigitte Jörg (DFKI, DE), Geert van Grootel (EWI, BE), Simon Lambert (STFC, UK), Elly Dijk (KNAW, NL), and Thibaut Levy (ESF, FR), all of whom provided contributed sessions within the conference theme at CODATA 2008 organized by the authors.

7 REFERENCES

Asserson, A, Jeffery, K.G, & Lopatenko, A (2002) CERIF: Past, Present and Future. In Adamczak, W & Nase, A (Eds): *Proceedings CRIS2002 6th International Conference on Current Research Information Systems*; Kassel University Press ISBN 3-0331146-844, pp 33-40.

Asserson, A & Jeffery, K.G. (2004) Research Output Publications and CRIS. In Nase, A. & van Grootel, G. (Eds.) *Proceedings CRIS2004 Conference*, Leuven University Press ISBN 90 5867 3839 May 2004 pp 29-40

Asserson, A; Jeffery, K.G (2005).; 'Research Output Publications and CRIS' The Grey Journal volume 1 number 1: Spring 2005 TextRelease/Greynet ISSN 1574-1796, pp 5-8.

Jeffery, K.G. (1999a) Knowledge, Information and Data, September 1999. Paper submitted to Director General of Research Councils proposing the programme that became e-Science. Retrieved from the WWW, April 6, 2010: http://www.semanticgrid.org/docs/KnowledgeInformationData/KnowledgeInformationData.html

Jeffery, K G. (1999b) An Architecture for Grey Literature in a R&D Context. *Proceedings GL'99* (Grey Literature) Conference, Washington DC, October 1999.

Jeffery, K G. (2002) Metadata. In Brinkkemper, J., Lindencrona, E., & Solvberg, A. (Eds.), *Information Systems Engineering*. Springer Verlag, London 2000. ISBN 1-85233-317-0.

Jeffery, K.G., Lopatenko, A., & Asserson, A. (2002) Comparative Study of Metadata for Scientific Information: The Place of CERIF in CRISs and Scientific Repositories. In Adamczak, W. & Nase, A. (Eds.), *Proceedings CRIS2002 6th International Conference on Current Research Information Systems.* Kassel University Press ISBN 3-0331146-844, pp 77-86.

Jeffery, K.G. (2004) The New Technologies: can CRISs Benefit. In Nase, A. & van Grootel, G. (Eds.) *Proceedings CRIS2004 Conference*. Leuven University Press ISBN 90 5867 3839, pp 77-88.

Jeffery, K.G. & Asserson, A. (2005) Relating Intellectual Property Products to the Corporate Context. *Research Publication Quarterly 21(1).* ISSN 1053-8801, pp 18-26 .

Jeffery, K.G. (2005a) CRISs, Architectures and CERIF. *CCLRC-RAL Technical Report RAL-TR*-2005-003 (2005)

Jeffery, K.G. (2005b) CRIS + Open Access = The Route to Research Knowledge on the GRID. Invited talk; *IFLA2005*, Oslo. Conference Proceedings Session 101 (in English, French, Russian).

Jeffery, K.G. (2007) Infrastructure and Policy Framework for Maximising the Benefits from Research Output. Keynote in Chan, L. & Martens, R. (Eds.), *Proceedings 11th International Conference on Electronic Publishing (ELPUB2007): Awareness, Discovery and Open Access*, Vienna, Austria. IRIS-ISIS Publications 2007 ISBN 978-3-85437-292-9, pp 1-12.

Jeffery, K.G. & Asserson, A. (2006a) CRIS Central Relating Information System, In Asserson, A. & Simons, E. (Eds.), Enabling Interaction and Quality: Beyond he Hanseatic League. *Proceedings 8th International Conference on*

Current Research Information Systems CRIS2006 Conference, Bergen. Leuven University Press ISBN 978 90 5867 536 1.

Jeffery, K.G. & Asserson, A. (2006b) Supporting the Research Process with a CRIS In Asserson, A. & Simons, E. (Eds.), Enabling Interaction and Quality: Beyond the Hanseatic League. *Proceedings 8th International Conference on Current Research Information Systems CRIS2006 Conference, Bergen.* Leuven University Press ISBN 978 90 5867 536 1, pp 121-130.

Jeffery, K.G. & Asserson, A. (2006) Grey in the R&D Process.The Grey Journal 2 (3). ISSN 1574-1796.

Jeffery, K.G. & Asserson, A. (2008) CRIS: Research Organisation View of the e-Infrastructure. In Bosnjak, A. & Stempfhuber, M. (Eds.), *Proceedings CRIS2008 Conference*, Maribor, Slovenia. ISBN 978-961-6133-38-8, pp149-158

Lopatenko, A. Asserson, A, & Jeffery,K.G. (2002) CERIF: Information Retrieval of Research Information in a Distributed Heterogeneous Environment. In Adamczak, W. & Nase, A. (Eds.), *Proceedings CRIS2002 6th International Conference on Current Research Information Systems*. Kassel University Press. ISBN 3-0331146-844, pp 59-68.

Zimmerman, E. & Jeffery, K.G. (2004) The Need for a Standardised Current Research Information System (CRIS): A Call to Arms. In Nase, A. & van Grootel, G. (Eds.), *Proceedings CRIS2004 Conference.* Leuven University Press. ISBN 90 5867 3839, pp 153-160.