An introduction to CERIF

Document details

Author:	Rosemary Russell, UKOLN, University of Bath
Date:	June 2011
Version:	1.0
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.

1 Introduction

This document provides a short introduction to the CERIF (Common European Research Information Format) standard. It is mainly aimed at a UK audience, in particular to support the Research Information Management (RIM) projects funded by JISC. It may also serve wider communities eg information and research managers seeking a preliminary briefing before accessing more detailed technical information.

CERIF is often described as a very complex standard. In an effort to keep this document as simple as possible (while still covering the key elements) many parts have been omitted. The euroCRIS website[1] provides access to a range of materials on CERIF and related issues, including eg the standard itself, tutorials and task group documentation[2]. Much of the growing body of project and implementation documentation is included in the references.

2 What is CERIF?

CERIF is a standard for managing and exchanging research data, ie information about researchers, organisations, projects, outputs and funding, arising from the research process. It provides a data model that can be used to describe the research domain, including relationships between the constituent parts, and how these change over time.

Officially CERIF is a European Union Recommendation to member states; it was originally developed with the support of the European Commission.

3 Background/History

CERIF originated in a European project (IDEAS) in the early 1980s which investigated linking databases of research information and was followed by EXIRPTS which extended involvement to the US and Japan. The end result was CERIF91, a simple standard based on records describing projects; persons, organisations and other information were represented as attributes. However, it was soon realised that CERIF91 was inadequate (too rigid in format, did not handle repeating groups of information, not multilingual etc). The subsequent CERIF2000 data model included results from projects, as well as organisations, persons, expertise, equipment and facilities. CERIF 2000 introduced three core entities: OrgUnit, Person and Project; relations between these were made using link entities. In addition CERIF was extended to accommodate Dublin Core, recognising the requirements of the Grey Literature community and the increasing number of institutional repositories based on OAI-PMH/DC[3]. In CERIF2006, ResultPublication (which already existed as an entity) was added to the existing three core entities; roles and types were also reorganised, to form the 'Semantic Layer'.

The current release is CERIF 2008 – 1.2 (November 2010), the final release in the 2008 series. It includes a major upgrade, providing formal CERIF Semantics for a defined, current core of entities. It also extended the publication entity, in response to requests for interoperability with institutional repositories. The next major CERIF release is planned during 2011.

4 euroCRIS

In 2000 euroCRIS became the official custodian of CERIF. euroCRIS is a not-for-profit association which aims to be the international point of reference for all matters relating to Current Research Information Systems (CRIS); it is not therefore restricted to supporting and promoting CERIF, although that is its primary task.

euroCRIS organises a biennial conference as well as biannual membership meetings (with tutorials) which aim to involve members in euroCRIS initiatives. There is also an annual seminar and joint workshops such as the 2nd CERIF CRIS and Institutional Repositories workshop which took place in May 2011. The core of the work is carried out in euroCRIS task groups; the CERIF task group is responsible for maintaining the standard and has regular meetings as well as a discussion list. The task group is responsive to requests for changes and additions to the standard, as demonstrated by the recent changes made to accommodate UK requirements.

A set of discussion fora have been set up on the euroCRIS website, covering CRIS architecture, best practice, CERIF, CRIS-Institutional Repositories and projects; these are restricted to members only. Members have access to additional resources via the website, such as draft releases and meeting reports;

they also receive a regular ‘Newsflash’ which provides useful updates on CERIF developments and implementations and other CRIS news, with items contributed by members. However it is recognised that more CERIF introductory resources are needed, to help with the initial steep learning curve.

Membership of euroCRIS is open to all those working with CRIS; members can joint at institutional, personal and affiliate levels. The UK now has the largest membership, with 30 institutional members at the time of writing (next are Germany, Finland and the Netherlands with 6 institutional members each). The majority of UK members have joined within the last year. Therefore much of this growth is a direct result of recent JISC funding of CERIF-based projects. CERIF-related activity in the UK effected modifications to the standard in 2010: the R4R and CRISPool projects requested changes (including classifications) on behalf of the UK community. These were approved by the CERIF Task Group and included in the CERIF 2008 – 1.2 release.

5 CERIF 2008-1.2 release components

The current CERIF 2008 – 1.2 release (November 2010) comprises the following components:

- CERIF 2008 – 1.2 FDM: Model Introduction and Specification
- CERIF 2008 – 1.2 FDM: SQL scripts for most common databases (members only)
- CERIF 2008 – 1.2 XML: Data Exchange Format Specification
- CERIF 2008 – 1.2 XML examples (members only)
- CERIF 2008 – 1.2 XML Schema files
- CERIF 2008 – 1.2 Semantics

All the above documents can be accessed via the euroCRIS website. This introduction focuses on the Model Introduction and Specification, and CERIF Semantics

6 CERIF 2008-1.2 conceptual structure

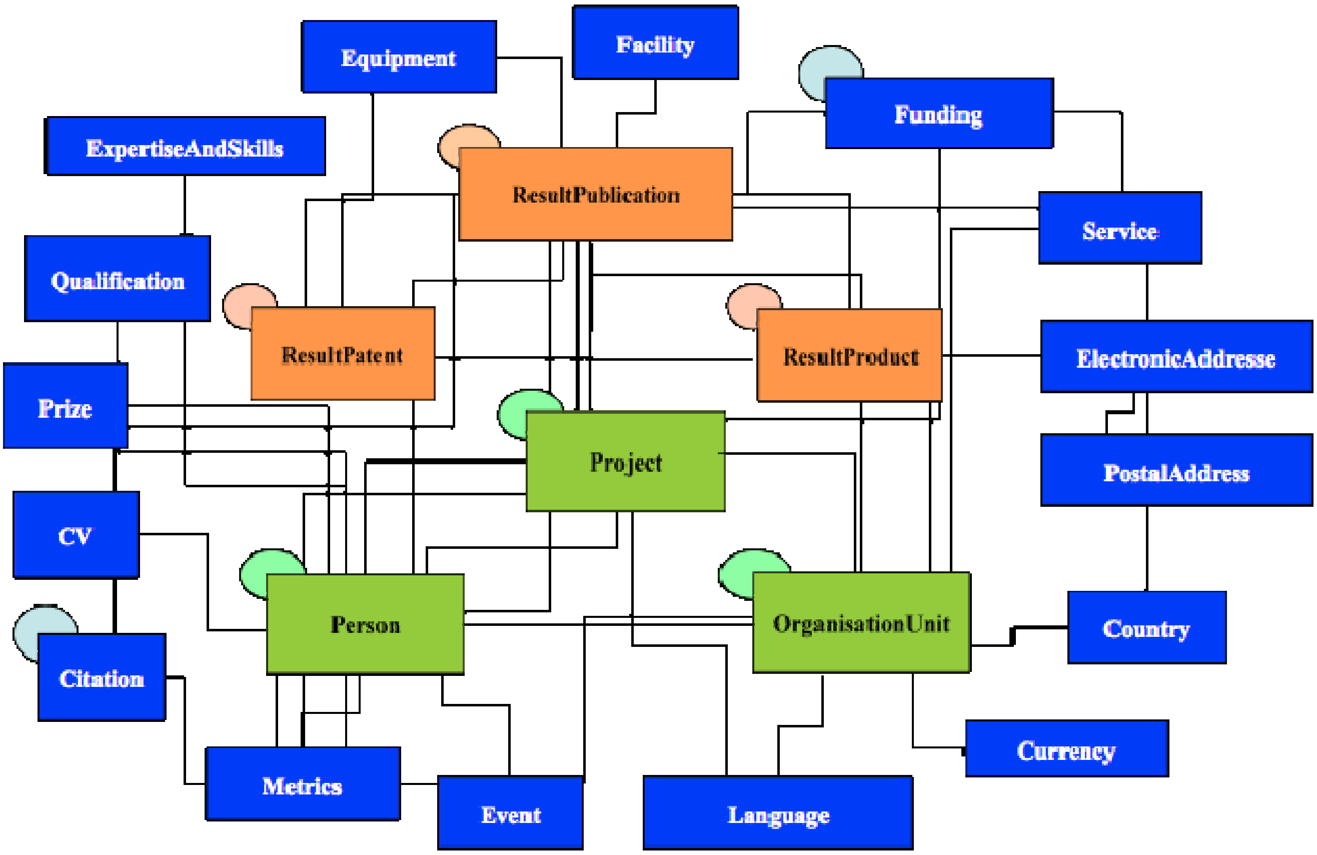

Figure 1: CERIF entities and their relationships[i]

[i] Reproduced with kind permission from euroCRIS

The conceptual structure of the CERIF data model provides a formal abstract description of the research entities (things being described) and their relationships. It is composed of entity types and features; these are explained below.

This section aims to provide an overview of the key entities in the model. It is closely based on the euroCRIS document *CERIF 2008-1.2 Full Data Model: Model Introduction and Specification*[4] and CERIF tutorials[5]. Full descriptions and many example records are provided in the specification so are not reproduced here.

6.1 Entity types

There are four types/levels of entity:

- Base entities (shown in green above)
- Result entities (orange)
- 2nd level entities (blue)
- Link entities

6.1.1 Base entities

Providing the key underlying structure for CERIF are the three base entities, as depicted in figure 1:

- Project
- Person
- OrganisationUnit

Entity names at conceptual/abstract level are presented in full, eg Project, whereas at implementation (physical) level they are abbreviated and given the prefix 'cf' for CERIF ie cfProj, cfPers, cfOrgUnit. This ensures the consistency of SQL scripts across databases by avoiding uncontrolled truncations (since many databases restrict the length of table names to a defined number of characters).

In figure 1 the ovoids attaching to each base entity indicate recursiveness (or self-referencing), meaning that relationships can be made within one entity eg cfProj_Proj, cfPers_Pers, cfOrgUnit_OrgUnit. Therefore hierarchical (1:n) or network (n:m) relationships could be recorded between persons or within an organisational unit.

As mentioned above, the project entity was the starting point for the CERIF standard, developed in order to exchange standardised information about research projects between EU member states. Person and OrgUnit were added as 1st level entities in CERIF2000, and in CERIF2008 were renamed as 'Base' entities.

CERIF lists Project attributes which are commonly used, including an identifier attribute (cfProjId), as well as acronym, URI, and start/end date (cfAcro, cfURI, cfStartDate, cfEndDate). Other attributes can also be added by first defining a new entity for the additional attributes.

Commonly used Person attributes listed are again firstly an identifier attribute (cfPersId), plus date of birth, gender and URI (cfBirthdate, cfGender, cfURI). CERIF enables maintenance of multiple person names or name variants using cfPersName and cfPersName_Pers.

Thirdly, OrganisationUnit has the identifier attribute cfOrgUnitId, in addition to acronym, currency, headcount, turnover and URI (cfCurrCode, cfAcro, cfHead, cfTurn, cfURI).

All the base entities can link to many other entities and can support multilingual features; this functionality is considered below.

The base entities therefore indicate/embody the ongoing central purpose of the CERIF standard, in recording data about projects, people and organisations associated with the research process, thus enabling organisations to provide a subsequent set of services based on standardised and reliable data.

6.1.2 Result entities

Closely associated with the base entities are the result entities:

- ResultPublication
- ResultPatent

- ResultProduct

As resulting outputs from the research process, publication is clearly the most commonly used entity of the three. It is also a ‘core’ entity - this is discussed further in section 7 below; result entities are however conceptually separate. Like the base entities, ResultPublication recursively links to itself.

In addition to the identifier attribute (cfResPublId), common result publication attributes are: publication date, number, volume, edition, series, issue, startpage, endpage, total pages, isbn, issn, and uri (cfResPublDate, cfNum, cfVolume, cfEdition, cfSeries, cfIssue, cfStartPage, cfEndpage, cfTotalPages, cfISBN, cfISSN, cfURI). Many relationships with other entities are maintained (as documented in the Full Data Model). The publication entity also supports multilingual features for title, subtitle, abstract, note, abbreviation and keywords.

ResultPatent and ResultProduct follow a similar pattern, with examples given in the Full Data Model document. The use of the patent entity is obvious. Product typically is the research dataset(s), software and any prototype products from a research activity.

6.1.3 Second level entities

Second level entities are shown in blue in figure 1, encircling the base and result entities. Second level entities enable representation of the research context (eg country, language, event, funding) via linking from the base and result entities. Minimum common attributes are identifier and URI. The funding entity includes eg FundID, CurrCode, StartDate, EndDate, Amount, URI.

6.1.4 Link entities

Link entities are the relationships or links between CERIF entities eg Person ‘is author of’ ResultPublication (see figure 2 below). Link entities are considered by euroCRIS to be a ‘major strength’ of the CERIF model. They are certainly integral to the entire structure. A link entity connects two entities, either base, result, or 2nd level entities eg ResultPublication ‘is funded by’ FundingProgramme (result entity linked to 2nd level entity) or one entity to itself (recursion) eg OrgUnit-OrgUnit..

Link entity identifiers are labelled as inherited because they do not originate in the link entities themselves, but are inherited from the base, result, or 2nd level entities. Each link entity carries semantics by reference to the CERIF Semantic Layer via the cfInheritedClassificationIdentifier and cfInheritedClassificationSchemeIdentifier. Each linking record also requires a startdate and enddate. The inherited identifiers and the date attributes form the basis of link entities.

Since link entities record the precise time-bound relationships between and within entities, a person might be a member of both a project and an organisational unit, for different periods of time. Clements points out that relationships can become extremely complex, and the strength of CERIF is that it is able to capture such web-like complexity with a data model that is essentially very simple.[6]

Figure 2 below shows some example link entities (indicated by arrows), with possible roles.

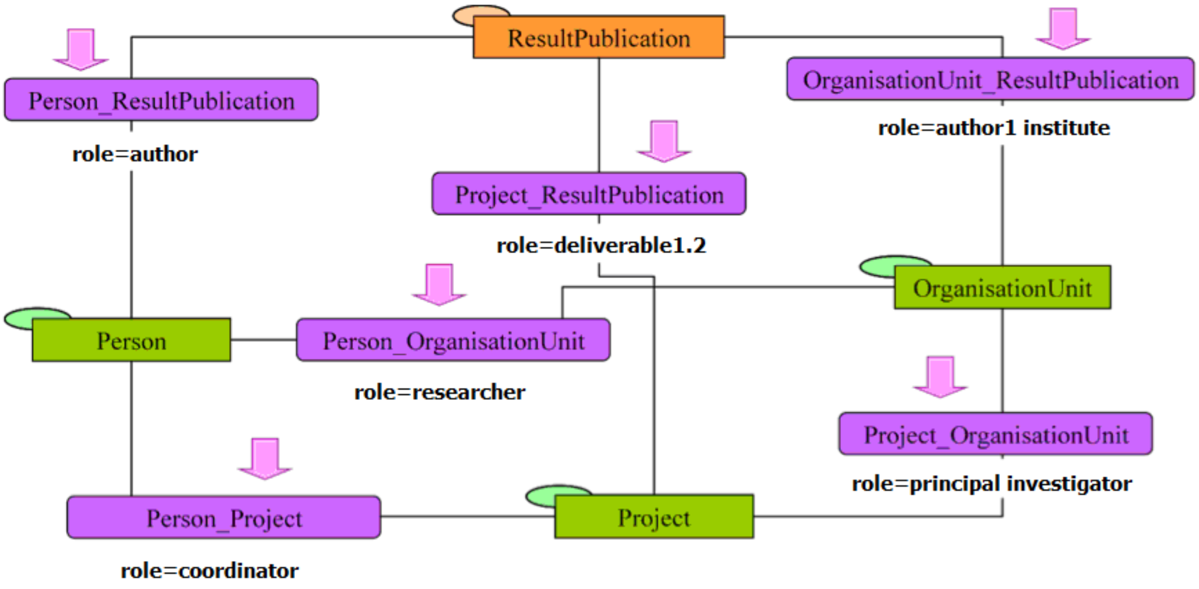

Figure 2: CERIF example link entities, with roles

6.2 Features

The second part of the CERIF conceptual structure is 'features':

- Multiple language
- Semantics
- Additional features

6.2.1 Multilingual features

The support of multilingual features is clearly very important within countries where several official languages are spoken and maintained. It is also important for collaborative research work, where the same project (or different workpackages within the same project) may be recorded in different languages by institutions in several countries working eg on EU funded research.

CERIF supports multiple language features for names, titles, descriptions, keywords, abstracts, and also for semantics (see next section). The encoded language is stored with the cfLangCode attribute that permits five character values eg en, de, fr, si, en-uk, en-us, fr-fr, fr-be, fr-nl.

The publication entity supports multilingual features for title, subtitle, abstract, note, abbreviation and keywords (cfResPublTitle, cfResPublSubtitle, cfResPublAbstr, cfResPublKeyw, cfResPublNameAbbrev).

6.2.2 Semantic features

The CERIF2006 release introduced the 'Semantic layer', which reorganised some of the existing semantic facilities (eg separating role and type definitions from entities). The semantic layer enables by the recording of relationship types, subject classifications, other classification schemes, simple mappings between schemas and thus semantic application views. It stores the semantic values used in attributes in base and secondary entities, as well as those in the link entities. This is done using the cfClassSchemeId attribute references, and it assigns each semantic value to a particular classification scheme. Therefore attributes in entities are assigned types and the relationships assigned roles (see above), in order to capture their semantics. Any schema or structure can be captured eg dictionaries, lexicons, thesauri, ontologies, and it is fully extensible. euroCRIS is considering supporting SKOS (Simple Knowledge Organisation System) in the future.

The CERIF semantic layer consists of the two class type entities: classification (cfClass), and classification scheme (cfClassScheme). Additionally it allows for a representation of multilingual terms (cfClassTerm) and class descriptions (cfClassDescr).

The CERIF Full Data Model describes the semantic layer as ‘a simple but powerful instrument’. The reorganisation succeeded in simplifying the data model and it certainly provides a lot of flexibility in capturing different application semantics and views, and allowing the assignment of multiple classification systems. However CERIF’s flexible approach means that it is not normally described as simple (see 8.1 below).

6.2.3 Additional features

The current CERIF release contains Dublin Core and Formalised Dublin Core entities and their attributes. In the 2011 releases euroCRIS plans to map from CERIF to Dublin Core, rather than keeping the Dublin Core elements within the physical model. The PersonName entity is currently also categorised as an additional feature, as it does not fit into the conceptual structure otherwise.

7 CERIF 2008 semantics

As a companion document to the Full Data Model, euroCRIS has also produced a separate semantic document: CERIF 2008 – 1.2 Semantics (see CERIF components in section 5 above). Whereas the Full Data Model presents the concept of the semantic layer within the CERIF data model, this document provides content – it presents a formalised collection of definitions for current core terms in their research context. euroCRIS defines the relationship as: “The CERIF Semantics may be best viewed as a filler; stuffing the CERIF Semantic Layer with contextually relevant semantics.”

The November 2010 release provided a major upgrade, with a large number of terms added by contributors to the content from many domains. Whereas the first version mainly focused on publications, version 1.2 presents a broader view (although publications terms have also been expanded, allowing complex classification).

As discussed above, CERIF 2008 has introduced five current CERIF ‘core’ entities:

- Project
- Person
- OrganisationUnit
- ResultPublication
- Funding

As indicated the core entities are not part of the conceptual CERIF model, but provide content for the CERIF semantic layer. euroCRIS states that the current core is only a temporary definition, so further entities may be included in future releases. Separating entities from the semantic layer allows greater flexibility when sharing local data with another institution.

Aims for future releases include improving descriptions, in order to disambiguate terms according to relationship contexts. The CERIF Task Group is working on the next draft.

8 CERIF in use

CERIF can be used in three ways:

- as a model for implementation of a standalone CRIS (but interoperation-ready)
- as a model to define the wrapper around a legacy non-CERIF CRIS to allow homogeneous access to heterogeneous systems
- as a definition of a data exchange format to create a common data warehouse from several CRIS.

There are many existing examples of each type of use. A useful set of CERIF CRIS case studies was produced for the DRIVER project by Mikael Karstensen Elbæk, covering Denmark, Ireland and the Netherlands[7]. Denmark, for example has had a centralised national CERIF CRIS system for many years. Trinity College Dublin and other universities in Ireland have been successful in integrating their CERIF CRIS and institutional repositories. However institutions with existing fit-for-purpose non-CERIF CRIS can still realise significant benefits by installing a CERIF ‘wrapper’ to act as an interface between their internal data structures and the external environment. Bolton estimates the cost of developing a CERIF wrapper at £13,000[8].

Some CERIF CRIS systems have been developed by institutions in-house (such as Trinity College Dublin), but it is now more common to purchase a commercial solution. PURE from Atira and Converis from Avedas are examples of commercial systems built on the CERIF data model and in use across Europe. A number of UK institutions have bought PURE and Converis systems, particularly in the last year and are in the process of implementation. Symplectic is another UK-based platform which is becoming CERIF-compliant. Bolton estimates that between 10-15% of UK HE institutions already have a CERIF-compliant CRIS[9]. A UKOLN document produced to support the JISC RIM Call for proposals in July 2011 reports on UK CERIF initiatives[10]. UKOLN also maintains a set of broader resources relating to research information management activities, particularly in the UK[11].

A CRIS would normally be implemented using a subset or superset of the full CERIF model, depending on the size and complexity of the requirement.

It is interesting that a paper (based on an early version of CERIF - 2002) by the former CCLRC (now merged and renamed STFC) demonstrates that the CERIF standard as 'a simple yet flexible structure' could be incorporated not only into CRIS applications but also into corporate-type information systems[12].

It is also notable that while implementing CERIF-compatible systems will deliver benefits to UK institutions, for the full benefits to be realised a range of organisations (eg HEFCE, the Research Councils, HESA) need to implement CERIF-compatible data collection mechanisms. The wider UK RIM environment is explored in a further UKOLN document[13].

There is also potential for CERIF to be used in tandem with other complementary approaches. The EXRI-UK (Exchanging Research Information in the UK)[14] report recommended reviewing the option of adopting a Linked Data approach using CERIF as the data model. A linked data semantic web can be generated from (and refreshed from) a CERIF-CRIS. The CERIF Task Group is taking forward work in this area.

8.1 Is CERIF too complex?

> *[The] CERIF standard is very complex, almost too complex for most users to fully understand.*[15]
>
> *CERIF has [a] steep learning curve… CERIF familiarization did cost time… Less human friendly… But needed…*[16]

CERIF is often criticised for its complexity. The similar quotations above are from different ends of the CERIF expertise spectrum. However it can be argued that complexity is a by-product of flexibility and choice of options. Implementors of standards invariably still want to be offered some opportunities for meeting specialist or local requirements.

The CRISPool project has pointed out that CERIF 'is extensible without prejudicing the core data model thus providing guaranteed interoperability at least at the core level but not precluding even richer communication'[17].

One of the key recommendations from EXRI-UK at the end of 2009 was that CERIF should be the basis for the exchange of research information in the UK. As a result, a range of successful CERIF activities have already started, in a variety of institution types (some funded by JISC). However as already indicated, much implementation work is still in the early stages. It is also outside the scope of this document to discuss the benefits to be gained from using CERIF. However it is worth highlighting that despite the frequent comments on complexity and the steep learning curve, the Bolton report found that CERIF CRIS could still be cost-effective for smaller institutions (with modest research portfolios) if their pre-existing research information management systems are poorly developed[18].

9 Conclusions…

Despite the acknowledged steep learning curve and implementation effort (in addition to software) costs involved, there are significant advantages to be gained from using a CERIF CRIS to achieve a standardised and interoperable managed research information environment. With the current increased implementation in the UK, there is growing expertise available to share both nationally and internationally (eg a thriving UK PURE user group is undertaking collaborative model development work). The momentum is likely to continue, with the funding of further JISC RIM projects aiming to widen community participation in CERIF CRIS and improve interoperability.

10 Acknowledgements

The author is grateful to euroCRIS for permission to use a range of CERIF documentation, figures and tutorials which have formed a basis for this introduction.

11 References

[1] euroCRIS: http://www.eurocris.org

[2] euroCRIS: http://www.eurocris.org

[3] Jeffery, K.G. An Architecture for Grey Literature in a R&D Context. Proceedings GL'99 (Grey Literature) Conference, Washington DC, October 1999. http://epubs.cclrc.ac.uk/

[4] CERIF 2008-1.2 Full Data Model: Model Introduction and Specification. http://www.eurocris.org/Uploads/Web%20pages/CERIF2008/Release_1.2/CERIF2008_1.2_FDM.pdf

[5] Jörg, B. CERIF tutorial held in Bologna, May 2011. http://www.eurocris.org/cerif/tutorial/

[6] Sheppard, N. Learning How to Play Nicely: Repositories and CRIS. Ariadne 64, July 2010. http://www.ariadne.ac.uk/issue64/wrn-repos-2010-05-rpt/

[7] Vernooy-Gerritsen, M. (ed.) Emerging Standards for Enhanced Publications and Repository Technology: Survey on Technology. Amsterdam University Press, 2009. http://dare.uva.nl/document/150752

[8] Bolton, S. The Business Case for the Adoption of a UK Standard for Research Information Interchange. Report to JISC. July 2010. http://ie-repository.jisc.ac.uk/487/

[9] Bolton op cit

[10] Russell, R. Research Information Management in the UK: Current Initiatives using CERIF. A supporting document for the JISC RIM3 Call, July 2011. http://www.ukoln.ac.uk/rim/dissemination/2011/rim-cerif-uk.pdf

[11] UKOLN Research Information Management: http://www.ukoln.ac.uk/rim/

[12] Grąbczewski, E. et al. A Corporate Data Repository For CCLRC Using CERIF. [n.d.] http://www.thesoundmanifesto.co.uk/papers/EG/CDR01.pdf

[13] Russell op cit

[14] Rogers, N. et al. Exchanging Research Information in the UK. EXRI-UK: a study funded by JISC. December 2009. http://ie-repository.jisc.ac.uk/448/

[15] Ready for REF CERIF Workshop. UoL Library blog, 24 March 2010. http://uollibraryblog.wordpress.com/2010/03/24/ready-for-ref-cerif-workshop-kings-march-2010/

[16] Van Grootel, G. FRIS: Flander Research Information Space. www.irpps.cnr.it/it/system/files/vanGrootel_09_05_2010_Rome.ppt

[17] CRISPool: http://www.st-andrews.ac.uk/crispool/background/cerif/

[18] Bolton op cit