CERIF: Past, Present and Future: An Overview

Anne Asserson, UiB, Keith G Jeffery, CLRC, Andrei Lopatenko, MU

Summary

CERIF (Common European Research Information Format) provides a canonical reference data model at both data and metadata levels. As such it is a model for the development of new CRISs (Current Research Information Systems) and a template both for data exchange between CRISs and for mediating access to multiple heterogeneous distributed CRISs. CERIF originated in 1988 but was based on earlier work in several European countries. The CERIF91 standard had some defects which became apparent in use. In 1997 a working group of the EC was set up to produce CERIF2000. This is a formal datamodel and thus unambiguously implementable. The CERIF Task Group of euroCRIS is working actively on implementations, lessons learned and improvements.

1. INTRODUCTION

This paper is organised as follows. CERIF (Common European Research Information Format) has a history described in Section 2. Section 3 describes how it is used currently as a datamodel for CRISs (Current Research Information Systems) in several variants in several countries and raises some issues. A Task Group of the euroCRIS organisation (www.eurocris.org) is now the custodian of CERIF ensuring its integrity, flexibility and usability. Section 4 outlines some of the developmental directions for CERIF and discusses the relationship to the original aims and objectives. Section 5 concludes.

2. CERIF PAST

2.1 CERIF91

CERIF (Official Journal 1991) has its origins in the late eighties arising from the Liaison Committee of Rectors' Conferences of Member States and parallel, independent, work by several national Research Councils especially in projects IDEAS (Jeffery et.al 1989) and EXIRPTS (Naldi et. al. 1992). A Group was formed to formulate CERIF91 (vanWoensel 1988a); (vanWoensel 1988b). Experience with CERIF91 led, in 1997, to the requirement for a new CERIF standard. The major aspects were:

a) the original CERIF covered only research projects as entities with persons, organisations and other information represented as attributes. Users of CRIS wanted to extend it to data on persons, organisations and other entities;

b) the "research subject classification scheme" recommended in CERIF 1991 had not been updated since 1988 and needed to be extended to cover the new data areas plus give enhanced coverage of existing ones;

c) new technologies, in particular, the introduction of the Internet and World Wide Web, had changed the nature of basic CRIS activities and opened new ways to serve various CRIS user groups.

The (CERIF2000) standard was created in late 1999 following two years of work by the Group formed to undertake this task. There were strong interactions with the ERGO (European Research Gateway Online) Group which was implementing a prototype portal system (ERGO) based on (a slightly extended) CERIF91.

2.2 CERIF2000

2.2.1 Problems with CERIF91

One of the major problems with CERIF91 – and operational CRISs from the eighties and nineties – was that they tended to have a single-entity focus. There were three main classes of systems:

(a) those focused on projects, with other information as attributes e.g. ASCENDA (UK);

(b) those focused on persons, with other information as attributes e.g. BEST (UK) or COS (USA);

(c) those focused on organisational unit, with other information as attributes e.g. LABO (FR);

the characteristics of all of them included:

(1) Problems of repeating groups. For example, in the case of CRISs focused on projects, it was not possible to record accurately the relationship between a person and this project. In fact usually only the project leader was recorded. Thus there were problems:

   (i) being unable to record repeating groups (multiple instances of groups of attributes representing e.g. person repeating against one project);

   (ii) having attributes with the same value (e.g. a group of attributes representing person) occurring multiply in the database – where the same person was associated with more than one project

   Similar problems occurred with repeating of organisational unit, publication, equipment, facility etc etc. This is known in the database theory literature as a problem of functional dependency.

(2) Problems of relationships. For example, in the case of CRISs focused on projects:

   (i) it was not possible to record that project A was a subproject of project B,

   (ii) nor a follow-on project from project B.

(iii) Similarly, it was not possible to record that person M was project leader, person N was the designer, person O was the analytical chemist for project A.

(iv) It was not possible to indicate that project A was a cooperation between two organisational units.

This all pointed to deficiencies in the data model. Specifically, it indicated that it was necessary to define more entities (rather than attributes of an entity) and that it was necessary to represent relationships between those entities that included 1:n, n:m and recursion (self-referencing).

2.2.2 CERIF2000 Design

(CERIF2000) has a particular feature of three major entities {project, person, organisational unit}interlinked through n:m relationships with {role} and {date / time} attributes and each capable of recursive reference (e.g. the relationship between one organisational unit and another, one project and another, one person and another). This provides great flexibility and robustness because not only can complex role and date-limited relationships between the three major entities be expressed but also other entities can be linked by role/date relationships to any or all of these three major entities. The following example facts can all be recorded accurately by CERIF2000:

(1) person *a* works for organisational unit *j* which is a sub-unit of organisational unit *k* which is a sub-unit of organisational unit *l*

(2) person *b* works for organisational unit *m,* a sub-unit of organisational unit *n*

(3) result-publication *x* came from project *p* which is a sub-project of project *q*

(4) person *c* is a reviewer of result-publication *x*

(5) person *d* is the editor of the journal or proceedings containing result-publication *x*

(6) organisational unit *h* (a publisher) claims copyright on result-publication *x*

(7) person *a* transferred copyright to organisational unit *h*

(8) for result-publication *x*, person *a* transferred copyright

and since all these statements include roles (eg author) and also date/time stamping {<start date/time><end date/time>}it is possible using, for example, date range intersection, to induce from (7) and (8) into person *a* transferred copyright to organisational unit *h* for result-publication *x*.

It is clear from the example that CERIF has both tremendous expressive power yet has the flexibility to allow simplified instances of the data model – for example in an academic environment the <date/time> attribute could be, simply, academic year thus allowing easy retrieval of result-publications of academic year yyyy with authors and organisational units (and

projects if desired). However, it does not even end there. Additional unique features of (CERIF2000) which provide even greater flexibility are:

(a) all contact information is stored in one entity which has relationships (with role and datestamping) to person and to organisational unit. Thus a person may have different contact information instances for different roles;

(b) all attributes with enumerated lists of valid values have those values stored in an entity with a relationship to the entities including the attribute thus providing flexibility and extensibility;

(c) all textual attributes have subordinate entities with language variants to allow a clean, flexible and extensible implementation of multilinguality;

(d) CERIF is delineated by key reference links to databases known to be pre-existing with more detailed information on certain entities e.g. publications, patents

2.2.3 Three Data Models

(CERIF2000) also proposes three data models:

(a) 'full CRIS' datamodel which defines entities, attributes and relationships for a 'greenfield' CRIS implementation;

(b) export CRIS datamodel which provides a set of proposed subsets of (a) for data exchange between CRISs capable of exporting / importing CERIF;

(c) CRIS metadata datamodel, a proper subset of (b), which provides a succinct description of the contents of a CRIS in a form readable by any CRIS capable of importing / exporting CERIF and also forms the key to the development of portal systems wishing to provide a homogeneous view over heterogeneous CRISs.

3. CERIF PRESENT

3.1 Introduction

CERIF is established. EC (European Commission) tenders in the area of ERIS (European Research Information System) emphasise CERIF. It is - with the CRIS Conference Series - a major raison d'etre of euroCRIS. It is used either in practice or as a best-practice reference. Utilisation of CERIF in practice has advanced our knowledge.

A list of current known CRISs which have CERIF compatibility is given:

SICRIS http://sicris.izum.si/default.asp?lang=eng : a CRIS providing access to total university research in Slovenia. It is highly CERIF-2000 compatible, based on MS SQL installation of CERIF-2000. Uses CERIF schema and CERIF vocabularies

AURIS-MM: The CRIS developed to provide access to Austrian university research. It is highly CERIF-2000 compatible, based on Oracle installation of CERIF-2000. It uses CERIF schema and CERIF vocabularies. It extends CERIF to serve better Austrian users (uses Austrian vocabularies), to deal with other information (multimedia, web sites) and to serve for better information retrieval (full-text indexes, views)

CRIS-MER http://www.ercomer.org/research/ReSchools/Re_plans.html : under development for research information on Migration and Ethnic Relations. Highly CERIF-2000 compatible, RDBMS implementation, uses CERIF schema and vocabularies. It extends CERIF for humantitarian information (new vocabularies and relations)

Scottish Research Information System http://www.scottishresearch.com : is a CRIS for public research in Scotland. It is CERIF-2000 compatible. The data schema and metadata schemas to describe data are based on CERIF-2000.

ARAMIS http://www.aramis-research.ch : a CRIS Intended to provide information to interested parties about research which is financed or carried out by the Federal Government in Switzerland. It has CERIF-2000 compatible data structures.

INTACCOMP http://www.intaccomp.ro/ : is a network of key data about Central Europe research projects sponsored with either national or international funds. The data schema, vocabularies and metadata schema for data exchange are based on CERIF-2000 recommendation (http://www.man.poznan.pl/ist/isthmus/programme/slides/goczyla/ISThmus-goczyla.PPT)

SAFARI http://safari.vr.se/ : a CRIS to provide information to Swedish academic research already available on the Internet. It is based on metadata technologies. The metadata schema is based Dublin Core and utilizes CERIF vocabularies to classify subjects.

Joint Electronic Submission (Je-S): a proposed system for electronic submission of grant applications (and hence into databases) of the 6 UK Research Councils has specified CERIF compatibility.

Experience has shown that CRIS developed for public research in Universities are commonly very compatible with CERIF-2000, even if they were developed without knowledge of CERIF-2000. Austrian examples are University of Salzburg, Technical University of Graz, University of Linz but similar examples are found in all countries. This is not surprising as CERIF2000 was defined utilising the experience of CRIS managers from all over Europe.

Particular implementations at UiB in Bergen (emphasising research results-publications) and in CLRC near Oxford (as a corporate data model for a R&D enterprise to drive business processes and provide R&D management information for decision support) have stress-tested the CERIF model and led to proposals for some extensions.

In parallel, detailed technical work on the EC-provided variants of CERIF schemas at the website (www.cordis.lu/cerif) by Andrei Lopatenko at TUW (Vienna University of Technology) has

indicated some deficiencies, and – interestingly - some variations from the model defined by the CERIF2000 Group. Andrei has also implemented a CERIF-compatible CRIS at TUW named AURIS-MM and has provided a 'clean' version of CERIF for euroCRIS.

Thus we have several kinds of CERIF-developments today:

(1) corrections to the EC-provided datamodel and schemas;

(2) extensions to CERIF for research results-publications;

(3) extensions to CERIF for corporate data model usage;

(4) precision of CERIF dictionaries (lists of valid terms) associated with attributes that have the property of an enumerated list of possible valid values;

These developments all aim either:

(a) to make precise and formal the definition of CERIF and to formalise its change control processes to ensure clarity;

(b) to extend the capability and usability of CERIF for supporting CRIS in the widest sense, with extensions both in depth (detail) and in breadth (business requirements areas supported);

One interesting feature is that of all the extensions to CERIF proposed very few actually require extensions to CERIF – the original datamodel had the capability to represent the requirement. This is a testament to the skill and ability of the CERIF2000 Group.

3.2 Extensions

3.2.1 UiB

UiB had a particular need to relate {result_publication} to {person a} who at the time was working for {orgunit n} and to {person b} who was working for {orgunit m} In other words, they wished to relate a particular {result_publication} to the intersection of {person} and {orgunit}.

This can be expressed in CERIF2000. However, it requires the induction that if the date range (with appropriate role) in the relationship {person-result_publication} intersects the date range in the relationship {person-orgunit} then the person was working for that {orgunit} at the time of publication. Of course, the {person} could have been working for more than one {orgunit} and more than one {person} could have been working for the same {orgunit} and on the same {result_publication}.

For reasons of efficiency UiB decided to implement this as ternary relation {person-orgunit-result_publication}, without role and dates and so constructing a specialised fixed ternary relationship. This has the advantage that induction is not required, and that there are fewer join operations during selection (search). It has the disadvantage that it is difficult to represent the role

and time relationship of a {person} to either a {result_publication} or to an {orgunit}, and it also makes it more difficult to handle multi-author publications (because of repetition of the other two key attributes in the ternary relation). In practice this has proved inefficient in implementation.

As an aside during this work it was noted that there is no (recursive or non-recursive) link table {result_publication-result_publication}. Such a link-table could be useful for handling semantics such as 'paper x in proceedings y' where the proceedings is clearly a separate publication or the relationship 'paper x is an extended journal paper from paper y given at conference z'. In (CERIF2000) the original idea as that publications were recorded outside of CERIF, and that CERIF should hold only a pointer (e.g. URI) to the publication. It is now clear that this is insufficient and thus we now propose that this feature is added to the CERIF2000 standard. It is completely in line with the philosophy of CERIF and is analogous to the recursive relationship of {project} or {orgunit}, or the non-recursive relationship between one {project} and another or one {orgunit} and another.

3.2.2 CLRC-RAL

CLRC has decided that it requires a corporate data model to underpin the drive to make all its business processes electronically supported. Work started on this independently of CERIF but after a relatively short time the model proposed was observed to be close to CERIF and so they were compared formally. The result was the adoption of CERIF but with extensions for this CLRC purpose. As CLRC is an organisation for the purpose of R&D it is perhaps not surprising that a CRIS data model should form a suitable basis for a corporate data model. However, CERIF was aimed originally at recording R&D information and not at supporting the business of R&D.

The major extensions required in the CLRC datamodel are as follows:

{project}: considerably more information including project plan, costs, milestones, deliverables;

{person}: considerably more information including annual performance assessment which itself includes work objectives and their achievement and learning and development needs and their achievement. The record of past positions within the organisation can be recorded in CERIF, as can an employee's manager and senior managers. Records of travel need to be added, related to project. The authority of one person over another can be recorded in CERIF but not the financial authority of a person (authorising expenditure by project). Although the CV of a person (as recorded within CERIF) can record skills or competencies, it is not necessarily in a form suitable for processing within the business of an organisation.

{orgunit}: CERIF does not provide for information on the mission or objectives of an orgunit, nor its terms of reference (e.g. for a committee). It does not provide for financial information of an orgunit (e.g. annual budget, invoices, orders) nor human resource aspects of an orgunit (how many staff-years of effort does it control).

Furthermore, the extension – a linking relation - to allow {result_publication-result_publication} noted above is required by CLRC. Current work is evaluating how much extension is required to {equipment} and {facilities} from the CERIF model to accommodate CLRC needs.

CLRC staff are still working on the details of the data model and expect to provide a full proposal for CERIF extensions for consideration by the CERIF Task Group of EuroCRIS (www.eurocris.org) in due course.

3.3 Precision and Formalisation

3.3.1 DataModel Corrections

Work while at the Technical University of Vienna by Andre Lopatenko has discovered errors in (CERIF2000), particularly in the EC-provided schemas driven from the extended entity-relation diagrams. A few inconsistencies were also discovered in the spreadsheet tables in the appendix of (CERIF2000). The current correct version of the datamodel is available within the documentation of the CERIF Task Group of EuroCRIS (www.eurocris.org/cerif).

3.3.2 Dictionaries, Thesauri and Ontologies

Given a formally correct datamodel (syntax) the next step to permit effective use of CRIS and effective data exchange or homogeneous access is converging the semantics (meaning). This requires agreed terms in dictionaries, thesauri or ontologies. Some work was done in this area using classification schemes (codes and meanings) and is documented (CERIF2000). However, many attributes were not subjected to rigorous content definition. Recent work by Andrei Lopatenko has provided a XML-encoded RDF description of CERIF which provides the basis for a definition of formal semantics, now being attempted using DAML + OIL. (W3C)

4. CERIF FUTURE

4.1 Introduction

CERIF clearly still has much to offer the CRIS designer, or the systems engineer providing import / export from a legacy CRIS to other systems. It provides a formalised reliable model. It appears – with the formal definition of the dictionaries, thesauri and ontologies - to be complete for CRIS requirements. Furthermore, it is clearly extensible for other purposes including a corporate business data model.

4.2 Data Exchange

CERIF can be used for data exchange with data from CRIS A being converted from CRIS A format to CERIF, transmitted to CRIS B, received as CERIF and stored in the format of CRIS B ready for use by users of CRIS B. Although it can be shown theoretically that this is accomplished easily, demonstration of this capability is a target.

4.3 Data Access

However, CERIF can also be used for access – that is provision of a portal to all CRISs for an end-user either attached to a particular CRIS or free-standing. The portal allows query expression in one expressive language and translation of that query to the target CRISs. They export their results as CERIF to be integrated at the portal for the end-user. The end-user then receives an answer consisting of the union of the results from the different target CRISs in CERIF format, ready for storing in the user's local CRIS or independently. Such a system is subject to access rights, copyright, IPR and other restrictions. Such a portal system has yet to be constructed, but the ERGO pilot demonstrated feasibility.

Here CERIF intersects with the (W3C) concepts of the 'semantic web' and the 'web of trust', both areas of active research by the authors among others.

5. CONCLUSION

CERIF has demonstrated the basic soundness of the datamodel both in formal correctness and in its designed-in flexibility. This provides optimism for its success in the future. Of the proposed extensions few have required changes to CERIF. Some extensions (UiB) provide, arguably, efficiency but at the expense of program maintenance effort. Others (CLRC) are required in order to utilise CERIF in a much wider environment (corporate business processes and information systems) than originally intended (CRIS).

Acknowledgements

The excellent work of the CERIF2000 Group is there for all to see. The authors acknowledge the work of colleagues particularly Johanne Revheim at UiB and Stuart Robinson at CLRC.

Contact Information

anne.asserson@ub.uib.no

keith.g.jeffery@rl.ac.uk

alopatenko@cs.man.ac.uk

References

CERIF2000 www.cordis.lu/cerif

ERGO www.cordis.lu/ergo

Jeffery,K; Lay,J; Miquel,J-F; Zardan,S; Naldi,F; Vannini-Parenti,I (1989) IDEAS: A System for International Data Exchange and Access for Science *Information Processing and Management* Volume 25 No 6 pp703-711, 1989.

Naldi, F; Jeffery, K; Bordogna, G; Lay, J; Vannini-Parenti, I A Distributed Architecture to Provide Uniform Access to Pre-Existing Independent, Heterogeneous Information Systems *RAL Report 92-003*

Official Journal (1991) Recommendation to the Member States to use the CERIF format In Official Journal of the European Communities, OJ L 189, 13th July 1991.

van Woensel, L (1988a) 'CERIF Manual' October 1988

van Woensel, L (1988b) Towards harmonisation of databases on research in progress – Final report of the European Working Group on Research Databases November 1988. Published by the Liaison Committee of Rectors' Conferences of Member States of the European Communities and Directorate General for Science, Research and Development of the Commission of the European Communities; financed by the Commission of the E.C., contract PSS*0058/B, compiled by Dr. L. Van Woensel.

W3C www.w3.org