THE CERIF MODEL AS THE CORE OF A RESEARCH ORGANIZATION

Keith Jeffery

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

ABSTRACT

A CERIF-CRIS consists of base entities with records describing components of the research and link entities describing relationships among records in the base entities. As an example, three base entities may contain records describing a person, a publication and a project while two link entities relate respectively the person to the publication in role author and the person to the project in role project leader. This powerful linking or inter-relating capability includes temporal as well as role aspects and inter-relates dynamically and flexibly all the components of R&D. The CERIF model can be extended to inter-relate appropriate information from legacy information systems in an organisation, such as those covering accounting, human resources, project management, assets, stock control, etc. A CERIF-CRIS can thus provide a flexible low-cost integration comparable with an ERP (Enterprise Resource Planning) System, particularly in an organisation with R&D as its primary business.

1 THE REQUIREMENT

1.1 Introduction

Researchers, historically, have tended to follow their own self-defined path of research based on curiosity, past learning and experience, and ability measured by success, usually in the form of peer recognition. Thus they have had personal, and often idiosyncratic, systems for managing their own research, based on various technologies from paper filing to personal computer hard disk filing. A 2006 survey at Council for the Central Laboratory of the Research Councils (CCLRC) now merged into Science and Technology Facilities Council (STFC) demonstrated that the majority of researchers did not hold personal copies (paper or electronic) of their own publications, particularly those prior to about 1990. They relied on the journal or conference proceedings stored in libraries. Similar problems occur with research datasets, software, and grey literature, such as technical reports and 'know-how' documents. When researchers work in teams, locally or, increasingly, globally, the problem is much greater.

R&D (Research and Development) is a creative activity. Historically it has flourished in a rather un-managed way. Systems have been in place for peer-review of research publication output for some hundreds of years. Experimental methods and observation methods have been formalized in different disciplines. Support of R&D by individual wealthy patrons, who supported Leonardo da Vinci for example, has generally been replaced by either commercial decisions on investment or funding of academic R&D based on assessment of research proposals by anonymous experts. However, in general, the active management of R&D is a recent phenomenon. It is only in the last 50 years or so that funding agencies (or commercial companies) have defined objectives for research programmes and attempted to encourage research along certain directions. The idea of measuring the output of research is even more recent, with the Thomson ISI system (ISI) for recording publications and measuring citations and impact being perhaps the best known although most modern evaluations of research output from a research group take into account many more factors including products, patents, success in obtaining funding, number of trained researchers produced, number and success of spin-out companies, etc.

There is thus a need for individual researchers, research groups, funding agencies, and commercial companies to have management systems for their research portfolio ranging over:

- research output: research output publications, datasets, software and research events;
- research support: facilities and / or equipment used;
- research know-how: technical methodological information;

- research funding: research proposals, research contracts;
- research output for technology transfer: products, patents;
- research management support: financial, organisational, human resource, and project management information with appropriate contacts (stakeholder relationships); and
- research information: access to global information on R&D in order to manage competition, cooperation, and evaluation.

The major requirement on R&D Systems for the researcher is that they should be researcher-centric: that is they should provide a view of everything of interest to the researcher in a structured manner that appears logical to the researcher in order to optimize the productive time of the researcher. In fact, the R&D system should be part of the 'researcher workbench' and sit alongside systems the researcher uses for finding publications, contacting collaborators, and submitting research proposals or travel claims.

The major requirement on R&D systems for the organisation is that they should provide the information required for decision-making to the benefit of the organisation. This may involve moving resources to support the best research or alternatively providing funds to support a department or group that is falling behind.

Selected views of the systems described above for researchers or organisations may be made available as information to others for purposes such as publicity, education (of scholars and of the general public), or offerings for technology transfer and commercialisation.

2 CERIF-CRIS

2.1 CRIS

CRIS (Current Research Information System(s)) have been a topic of interest for many years. However, in 1991 Jostein Hauge of the University of Bergen drew together experts in the topic from all over Europe and beyond to share experience and discuss the issues. This first CRIS conference is perhaps the starting point of the formal use of the term CRIS. EuroCRIS (www.eurocris.org) defined CRIS as 'A Current Research Information System, commonly known as "CRIS", is any information tool dedicated to provide access to and disseminate research information. This includes People, Projects, Organizations, Results (publications, patents and products), Facilities, and Equipment, …'

There are several kinds of CRIS, and they may be classified along two axes: purpose and technology. Along the purpose axis, CRISs may exist to manage projects, to advertise expertise, to record research output publications, or to provide an overlay of any of the above. Along the technology axis, they may use information retrieval technology, hypermedia (including WWW), structured (hierarchic, network or relational) database technology, or some combination.

2.2 CERIF

The CERIF (Common European Research Information Format) standard (CERIF), technically an EU recommendation to member states, arose from a recommendation of the Conference of European University Rectors and in parallel a recommendation of the heads of research funding organisations of the G7 countries. The initial 1991 standard was based on one record per project with researcher, organization, and other information as (repeatable) attributes. The information was semi-structured. For reasons concerned with lack of formality and precision in the format and the classification scheme proposed, the 1991 standard proved unattractive. In 1997 the EC convened a new group to update CERIF: the result in 1999 was the CERIF2000 standard, since entrusted to the care of EuroCRIS and improved / updated continuously.

The process to derive CERIF2000 was interesting. The group all contributed information on existing CRIS schemas and desirable schemas. By a process of formal reduction, integration, and verification, the group reached a formal extended entity-relationship model that was documented formally by the authors. In order to prove the model, a prototype database was implemented in Microsoft ACCESS and various queries supplied by the group used to test whether the model could provide answers. Thus the group had great confidence in the CERIF2000 data model.

CERIF consists of several types of entities, realised as relational tables with columns for attributes and rows for instances. The data structure is normalised to avoid data replication and thus avoids consequent problems for update integrity. Thus, complete information about a research project may be spread over many entities. The types of entities are: base tables, secondary base tables, link tables, language field tables, and lookup or code tables. Lookup tables provide a list of valid values for an attribute and are used in validation and intelligent input. Language-field tables provide attribute values in different languages (e.g., project title in English and French). The major information is stored in the base tables (and secondary base tables), and the data structure is recorded in the link tables.

2.3 CERIF Characteristics

The design of the CERIF Datamodel has some important characteristics. It is designed so that it can be extended (by adding new base entities and then link entities to integrate with the structure) while preserving backward continuity in the original structure to allow guaranteed interoperation between CERIF-CRIS. It can link to any other system using the link entities. It is normalised to avoid replication of data (and consequent update integrity problems) and to improve performance. The data model can be implemented using any technology from hypermedia to information retrieval (semi-structured) and on to knowledge-based systems. The data model follows formally first order logic and so is available for deduction and induction leading to greater potential utilization of the data. The data model also includes lookup tables (code or classification tables) allowing improved data integrity by validation at input/update time and permits intelligent user interfaces to utilise the information to provide user assistance.

The key to the design is the separation of base entities from link entities. The base entities, once populated, are rarely amended but may be appended with new information. The link entities are where the main update activity takes place because they record new relationships between records in the base entities. These new relationships may be input or they may be generated by deduction or induction.

Consider the following case illustrated in Figure 1: A person A is an employee of organisation O and a member of organisations M and N, both of which are parts of O. She is author of X in which O claims the IPR (intellectual property right) and project leader of P. In CERIF the following records would be in base tables: Person: A; OrgUnit: O,M,N; Publication: X; Project: P. The link tables would be: Person-OrgUnit: A-employee-O, A-member-M, A-member-N; OrgUnit-OrgUnit: M-partof-O; N-partof-O; Person-Publication: A-author-X; OrgUnit-Publication: O-IPR-X; and Person-Project: A-projectleader-P. In fact, the link tables include, as well as role, the temporal information concerning start and end date-time. In this example, it may be that when A authored X she was no longer a member of M. This relatively simple example illustrates the power of CERIF as a data model; the authors know of no other existing CRIS data model that can express it.

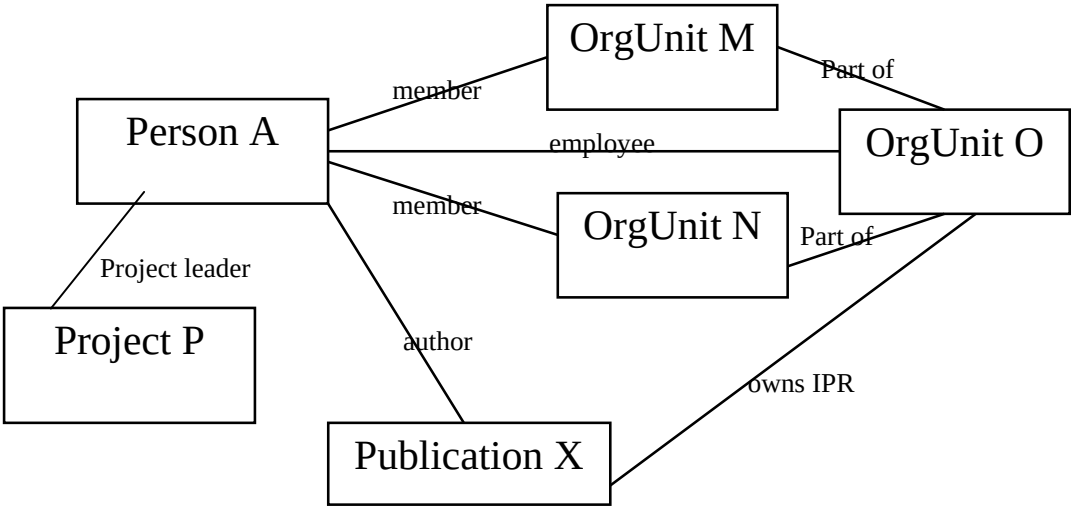

Figure 1: Example of CERIF Expressivity

Staying with this example, the linking to other systems' capability can be illustrated. The publication X full-text (or multimedia) is not stored within the CERIF data model but in an institutional repository or publisher's online database. CERIF provides the direct linkage to the full text. Similarly (subject to privacy and security) more information about Person A may be found in the HR (human resources) system of OrgUnit O or on web-pages associated with either OrgUnit M or N. Again, subject to privacy and security, the full project management information associated with Project P may be accessed in the project management system of Organisation O, and from thence, financial information may be found in the financial systems of Organisation O.

3 MANAGEMENT USING A CERIF-CRIS AS THE INTEGRATING SYSTEM

3.1 Current State

In most organisations engaged in R&D, the situation of management information is chaotic. At one extreme, individual researchers or research groups author web pages using HTML describing themselves, their projects, and their output publications. These web pages commonly become out-of-date as interest and enthusiasm wanes. At the other extreme, the finance system of the organisation records that a research grant has been obtained but has no record of what it concerns, who is involved within the organisation, with whom (if anyone) they are cooperating, and whether any output has been produced. Somewhere between a project management system (usually not covering the whole organisation but different systems in different groups) records that the research project has deliverables, milestones, and the state of their achievements. It may record resources planned and consumed.

The problem with this current state as described is that there is R&D information which is replicated, inconsistently and incompletely across the three example systems (and in fact there are usually many more systems involved). The problem reflects two underlying problems:

1. the traditional divide between the individual researcher or research group view of the world and the organisation management view of the world and
2. the traditional fierce independence of researchers and unwillingness to provide information on their activity. The first is characterised by the organisation wishing to ensure governance and value for money while the researcher wishes only to achieve peer recognition; the second is characterized by a quest for curiosity-led academic research freedom at any cost, despite possible advantages in cooperating with the management of an organisation. It may be compounded by the researchers' view that the IT (Information Technology) system they are using is inadequate and they could have designed it better!

3.2 Using CERIF

As indicated in section 2.3, CERIF not only provides a data model for recording the R&D information of an organisation, it also provides 'hooks' to link up to other systems. Let us consider those 'hooks' and how they may be used.

CERIF provides attributes for basic information about a person, not least to permit disambiguation and authenticated identification. CERIF also provides for a CV to be stored. However, it is likely that in an environment where more information about a person is required, access to the HR system of the organisation (subject to privacy and security) would be required. This can be provided by a link table entry for each person relating the value of the primary key of person in CERIF to that in the HR system.

In a way similar to person, CERIF records only basic information about an organisation. A link table could relate the primary key value of an organisation in CERIF to that in a system cataloguing organisations, or the internal organisational database reflecting its structure, or even the URL of a web-page describing the organisational unit.

Again, using the technique described above, the primary key value of a project in CERIF may be related to that in a project management system within the organisation. Alternatively, or better in addition, it could be related to the primary key value for that project in the database of the funding organisation allowing direct linkage to that information.

The primary key of the funding in CERIF can be related by link tables internally to OrgUnit (the funding agency) but additional link tables can relate it to, within the funding organization, a particular programme or other initiative.

CERIF provides contact information for an instance related through a link table to another base entity instance such a Person or OrgUnit. However, further contact information about an organisation or a person may be required. This information may be stored in a CRM (Customer Relationship Management) System, and using the usual technique, the instance of the contact within CERIF may be related to that within the CRM.

As indicated already above, the information concerning a publication in CERIF can be related via a link table to the full text or multimedia stored in an institutional repository and/or in the database of a publisher.

Full details on a patent are stored in a patent database, either national or European or both. The Patent information in CERIF may be linked to the information available in those databases.

There are many kinds of research output products. Common kinds include research datasets and software. These are commonly stored in an appropriate filestore system with a portal for access. Clearly, as with any other external system, a link-table in a CERIF CRIS can be used to relate the research output product to any CERIF entity such as project, person, or organizational unit or, indeed, publication(s).

We believe the above illustrates the flexibility of CERIF with respect to linkage; the same technique applies for other CERIF entities such as events, facilities, equipment, etc.

3.3 The Whole Picture

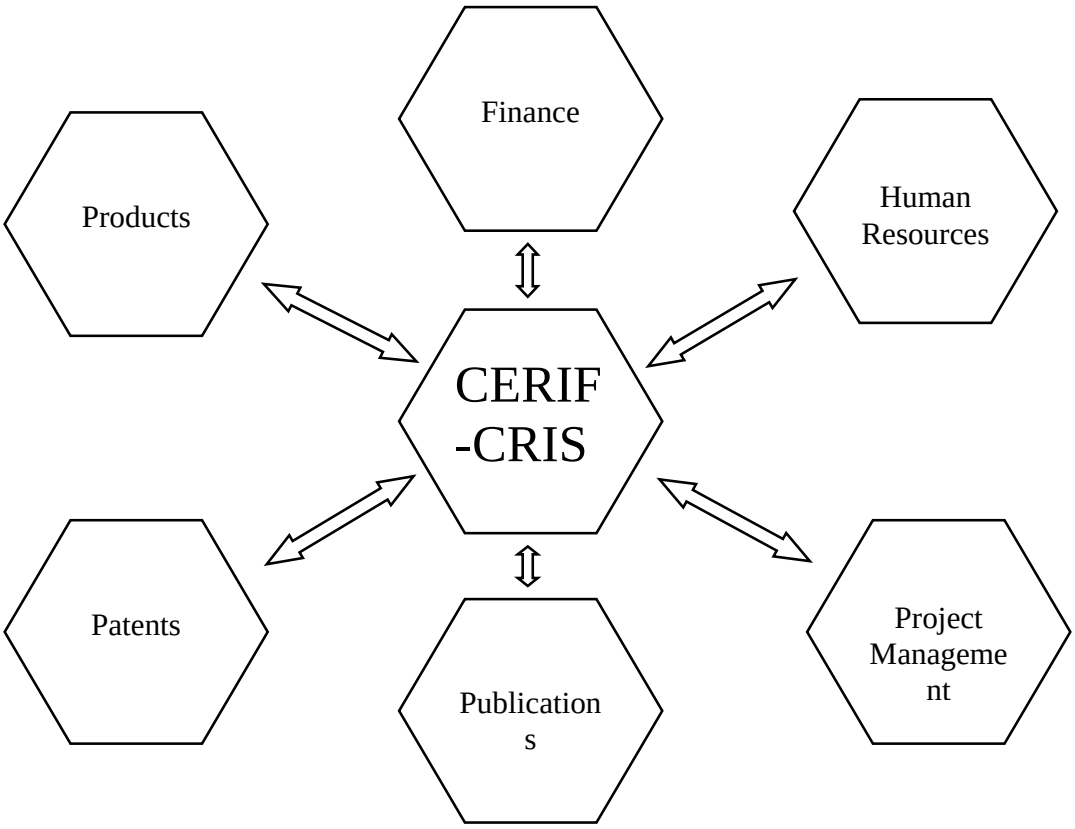

Figure 1: CRIS at the Centre

For the individual researcher, access to other research output information may be more important. For example, having found a relevant publication, the researcher may wish to access the research datasets and / or software used