16.1 Introduction

The last 10 years have witnessed a revolution in the research environment, partly mirrored in the commercial and social environments. The underlying factors concern the increasing price-performance of computer hardware including processing, storage and networks; the improvements in user interface technology including mobile phones making the ICT environment more readily available and - of course – WWW (World Wide Web). Because of the challenges in speeds, volumes and complexity, the research environment tends to anticipate by some years developments in the other environments. The e-Science concept, developed in UK from an initial paper by Keith Jeffery [Je99] encompasses and assumes an e-infrastructure [e-IRG] (in USA cyberinfrastructure [NSFCyb]) consisting of networks, computational servers, data servers and detectors. The e-Science concept, however, builds on this physical layer two more layers; one managing information (derived from data by structuring in context) and surmounted by a knowledge layer recording human-generated knowledge (such as scholarly publications) or computer-generated knowledge (derived through data mining).

Synchronously with e-Science, Anne Asserson, Keith Jeffery and others promoted the concept of CRIS (Current Research Information Systems) and the CERIF (Common European Research Information Format) EU recommendation to member states. CERIF is a rich and flexible data model for CRIS or for interoperation of CRIS with formal syntax and declared semantics – thus making it machine-understandable as well as machine-readable. However, CRIS are a necessary component of e-Science allowing researchers, research managers, educators, entrepreneurs and the media to discover what research is being done, by whom, in which organisations, through which projects, from where the funding comes and what are the outputs including publications, products and patents. Clearly, CRIS form an essential way in the e-research environment - including the e-infrastructure - to index research and make it available. It is common for a CRIS to be associated with a repository of full text (or hypermedia) objects such as scholarly publications i.e. one output of the research. However, the repository
equally may contain grey material such as technical reports which, in fact, may form a large component of the ‘know-how’ and IP (intellectual property) of an organisation.

This paper argues that the future of grey literature (in the widest sense) lies within the context of an e-research environment populated with CERIF-CRIS and associated repositories.

16.2 The e-Research Environment

In 1998-1999 the UK Research Council community was proposing future programmes for R&D. The author was asked to propose an integrating IT architecture [Je99a]. The proposal was based on concepts including distributed computing, metacomputing, metadata, agent- and broker-based middleware, client-server migrating to three-layer and then peer-to-peer architectures and integrated knowledge-based assist. The novelty lay in the integration of various techniques into one architectural framework [Je04].

The UK Research Council community of researchers was facing several IT-based problems. Their ambitions for scientific discovery included post-genomic discoveries, climate change understanding, oceanographic studies, environmental pollution monitoring and modelling, precise materials science, studies of combustion processes, advanced engineering, pharmaceutical design, and particle physics data handling and simulation. They needed more processor power, more data storage capacity, better analysis and visualisation – all supported by easy-to-use tools controlled through an intuitive user interface.

Figure 1: GRIDs architecture
On the other hand, much of commercial ICT (Information and Communication Technology) including process plant control, management information and decision support systems, IT-assisted business processes and their re-engineering, entertainment and media systems and diagnosis support systems all require ever-increasing computational power and expedited information access, ideally through a uniform system providing a seamless information and computation landscape to the end-user. Thus there is a large potential market for GRIDs systems to provide the e-Science (or more broadly e-Research) environment.

The original proposal based the academic development of the GRIDs architecture and facilities on scientific challenging applications, then involving IT companies as the middleware stabilised to produce products which in turn could be taken up by the commercial world. During 2000 the UK e-Science programme was elaborated with funding starting in April 2001.

The architecture proposed consists of three layers (Figure 1). The computation / data grid has supercomputers, large servers, massive data storage facilities and specialised devices and facilities (e.g. for VR (Virtual Reality)) all linked by high-speed networking and forms the lowest layer. The main functions include compute load sharing / algorithm partitioning, resolution of data source addresses, security, replication and message rerouting. This layer also provides connectivity to detectors and instruments. The information grid is superimposed on the computation / data grid and resolves homogeneous access to heterogeneous information sources mainly through the use of metadata and middleware. Finally, the uppermost layer is the knowledge grid that utilises knowledge discovery in database technology to generate knowledge and also allows for representation of knowledge through peer-reviewed scholarly works (publications) and grey literature, especially hyper-linked to information and data to sustain the assertions in the knowledge.

The concept is based on the idea of a uniform landscape within the GRIDs domain, the complexity of which is masked by easy-to-use interfaces. The achievement of this virtualisation is based on metadata [Je00] used in this context [Je04].

16.3 CRIS

CRIS have existed for many decades in research funding organisations and in some research performing institutions. However, it was not until 1991 that experience was shared internationally, although there had been initiatives to interoperate a limited number of CRIS as early as 1984. Driven by various pressure groups, the EC (European Commission) drew together a group of national experts in 1987-1989 to produce the first CERIF (Common European Research Information Format) recommendation. The expert group was reconvened in 1997 to produce the much-improved CERIF2000 recommendation upon which all subsequent development is based. In 2002 the EC requested euroCRIS (www.eurocris.org) to take responsibility for the promotion, maintenance and development of CERIF.
Full details of CERIF are available at www.eurocris.org/cerif. The original purpose of CERIF was to provide a data model for anyone developing a new CRIS and to provide a data model for interoperation between pre-existing (legacy) CRIS (Figure 2). CERIF was developed as a generic datamodel using advanced concepts [AsJeLo2002]. However, CERIF has also been used as a central directory system for an organisation and can be extended further to integrate legacy systems within an organisation [JeAs2006]. CERIF is now becoming more widely used in organisations engaged in R&D whether funders, policymakers, innovators/entrepreneurs, media or academic (research-performing) institutions.

![Figure 2: CERIF datamodel](image)

CERIF has some features worth highlighting because of their relevance not only in the research domain but much more widely.

First, CERIF assumes not a hierarchic model of the world but a fully connected (possibly cyclic) graph. This provides great fidelity in representation. For example, many systems have a hierarchic relationship between university department and academic staff member. CERIF can represent accurately an academic staff member related to multiple departments, multiple research groups, multiple academic institutions and commercial organisations.

Second, CERIF separates base relations - as fundamental entities of interest - from relationships. Thus, CERIF has the concept of person (as opposed to researcher, author, employee…) and the role of that person is defined in the relationship of that person to another entity such as an organisation (person P is em-
ployee of Organisation O) or to a publication (person P is author of publication X) or to another person (person P is co-author with person Q) (Figure 3).

Third, CERIF provides a ‘time machine’. This is done not by recording the valid time and transaction time on the base entity instances (the conventional temporal database approach) but by recording date-time-start and date-time-end on the relationship between instances of entities (person P start 20000801:09:00:00 end 20081231:17:00:00 is employee of organisation O). This means it is possible to re-create the history of an instance of an entity (e.g. the CV of a person) or to recreate the state of an organisation (persons, organisational structure, funding, outputs…) at a given date-time or period of time between start datetime and end datetime.

Fourth, CERIF is defined to use Unicode - so that any character set can be represented - and allows declaration of one or more languages for any textual attribute value. This means that multilinguality is handled effectively.

Fifth, CERIF – because of its first-order-logic structure, allows deduction and induction to generate new facts. This is important in saving unnecessary end-user input and permits the support of knowledge-assisted user input and validation.

16.4 Repositories

Repositories (of scholarly material) have developed within a library environment and mainly to record the output of research at an institution – the institutional repository – containing author-deposited copies of peer-reviewed published material; so-called green open access material. This contrasts with gold open access where the author institution pays for a publisher to make the material available under open access on the publisher repository system.
Repositories store and provide access to the detailed information. It is usual—and best practice—to separate repositories of research publications from repositories of research datasets and software (e-Science or, better, e-Research repositories) because of their different access patterns and different metadata requirements. The e-Research repositories require much more detailed metadata to control utilisation of the software and datasets in addition to metadata to allow discovery of the resources. At present they tend to be specific to an individual organisation because of their novelty and the differing requirements on metadata imposed by different (commonly international) communities e.g. in space science, atmospheric physics, materials science, particle physics, humanities or social science.

Publication repositories need not be restricted to peer-reviewed published material. Increasingly institutional repositories include e-preprints and technical reports i.e. grey literature. Some, indeed, include more informal material and teaching material, presentations and lecture notes. Publication repositories typically use some form of Dublin Core Metadata [DC] and most are [OAI-PMH] (Open Archive Initiative – Protocol for Metadata Harvesting) compliant for interoperation and are indexed by Google Scholar. Example software systems are [ePrints], [DSpace], [Fedora] and [ePubs]. Although the metadata associated with the publication includes author name, different publishers/journals/conference proceedings require the name to be in different formats so correlation—and disambiguation from other authors with similar names—is very difficult.

The publication or its metadata may contain information on the institution of the author, but usually only one such organisation even if the author is associated with multiple organisations. Information on the project from which the publication was generated, funding source, facilities or equipment used etc. may or may not be recorded within the publication but not in a structured form and so is more-or-less impossible to extract automatically. Publication repositories require the author to input metadata to describe the publication; this is a threshold barrier and can be reduced by utilising pre-recorded information in the CRIS. The combination of (a) the difficulty of extracting contextual metadata on research as described above from repositories and (b) the threshold barrier caused by human input of metadata leads inevitably to the conclusion that we should link together CRIS and repositories to gain the advantages of each.

Thus, there is an advantage in linking together repositories (with the full text or hypermedia publication and/or repositories with research data and software) with a CRIS which provides structured information on the context of the research—project, equipment, funding, organisations and persons involved [AsJe04]. The metadata in the CRIS describing scholarly publications may be used for evaluation of research; a well-known example is the Norwegian FRIDA [FRIDA] system.
16.5 Organisational ICT

Research funding organisations and research performing institutions need to manage the research. At present most institutions have a complex mix of legacy systems covering this requirement. Worse, commonly they have multiple protocols for intercommunicating with other institutions; an example is the submission of a research proposal from a university to a funding organisation and subsequent transactions involving research products and funding. A CERIF-CRIS can be used as the unifying system [JeAs2006a] over these legacy systems allowing an institution to continue to utilise legacy systems and to replace them as and when business conditions permit. A strong advantage of such a unifying CRIS is that it can be used to support both the workflow of organisational administrative processes and the entry of metadata. The latter involves ‘pre-completing’ web forms using information stored in the CRIS such as person name, organisation, contact information. Taking the case of a publication, commonly it starts life as grey literature and can be recorded in the CRIS (metadata) and in the repository (full text or hypermedia); if/when it becomes white literature the only additional metadata required concerns the bibliographic information of the publication channel – the remaining metadata information is already stored in the CRIS and thus can be reused [JeAs06b].

16.6 Interoperation

In addition to unifying the IT support of one organisation, the CERIF-CRIS can also be used to interoperate with other institutions thus supporting the distributed and international scale of research – or any commercial / industrial business or social activity. However, interoperation requires a common data format to reduce the many (n*(n-1)) interconversions (between every pair of nodes) to n (each node converts only once to the common standard). There are several architectures to achieve this which were described, characterised and compared [Je05], [JeAs08]: Remote Wrapper; Local Wrapper; Catalog; Catalog plus Pull; Full CERIF; Harvesting. Each has advantages and disadvantages although – obviously – the greatest benefits are obtained by interoperating fully-compliant CERIF-CRIS.

Nonetheless, organisations with legacy systems that are not CERIF-CRIS can utilise one of the techniques mentioned above to ‘wrap’ their existing system(s) so that interoperation / intercommunication with other organisations utilises CERIF as the canonical information exchange format. Indeed, a special group set up by ESF (European Sciences Foundation) at the request of euroHORCS (European Heads of research Councils) reached the same conclusions although euroHORCs decided it was too early for such interoperating systems but encouraged members to converge towards the architecture proposed.

The advantage of interoperating CRIS is that a researcher, research manager, innovator or media reporter can query in a homogeneous way across heterogeneous...
ous distributed research information sources – including onward access to repositories including peer-reviewed publications and grey literature. It makes possible answers to queries such as ‘which researchers are working on drugs to combat HIV/AIDS – sort by country and within country by institution’ or ‘how many peer-reviewed publications were produced between 1995 and 2000 on global warming – sort by country and within country by aggregated publication impact factor’. In each case the additional access to the repositories via the CRIS used as metadata can provide the full text or hypermedia publication.

The benefits are obvious. Researchers can find teams also working in their field – and this is especially important in emerging multidisciplinary fields where the existing subject- or specialism-based academic networks do not yet extend. Research managers can decide on strategy to compete or cooperate with other institutions or – at national scale – with other countries. Innovators can find research ideas relevant to their commercial interests. The media can find ‘science stories’ that popularise research with the general public and which can stimulate debate on pressing issues – including ethical and funding priority issues – in research; examples include discussions on global warming, GM (genetically modified) food, defence-related research etc.

16.7 Grey in Context

Grey literature – in the widest sense including hypermedia – is produced in the research process and provides a valuable resource. Indeed, commonly it forms the IP of an organisation (technological ideas described) and leads to innovation and wealth creation [JeAs04], or to improved effectiveness and efficiency (‘how to’ manuals) in the operations of the organisation.

Currently grey literature is usually stored in repositories. There is no common agreement on the metadata to be used to describe this resource although some organisations are using a version of DC and SIGLE [SIGLE} has a defined metadata standard. As suggested [Je99] grey literature (and also white literature) requires richer metadata - than that provided by DC - that has both formal syntax (for efficient computer processing) and declared semantics (to automate processes that would otherwise be performed by humans thus increasing effectiveness and efficiency).

The conclusion was that the metadata should be CERIF-compatible and stored in the CRIS, with the grey literature object – full text or hypermedia – stored in a repository with the two sources linked to allow optimal use of the characteristics of the CRIS and the repository [JeAs05]. In this way not only is the grey literature object provided with better metadata for retrieval but also is associated with the other contextual metadata in the CRIS covering projects, persons, organisations, facilities, equipment, events, products and patents. This truly puts ‘grey in context’.
With the CRIS forming the research context backbone information in the e-infrastructure supporting GRIDs and e-research, this further places grey firmly in the research environment together with other publications and products. This architectural approach positions optimally grey literature.

References


[CERIF] Commpn European Research Information Format page within euroCRIS website http://www.eurocris.org/ cerif


[DSpace] DSpace repository product homepage http://www.dspace.org/


[ePubs] ePubs reflection group roadmap: http://epubs.cclrc.ac.uk/

[eprints] ePrints repository product homepage http://www.eprints.org/

[Fedora] Fedora repository product homepage http://www.fedora.info/

[FRIDA] Norwegian FRIDA CRIS Homepage http://frida.uio.no

[Je99a] Jeffery, K.G ‘Knowledge, Information and Data’ September 1999 Paper submitted to Director General of Research Councils proposing the programme that became e-Science; available at http://www.semanticgrid.org/docs/KnowledgeInformationData/KnowledgeInformationData.html


[Je04] Jeffery, K.G.; ‘GRIDs, Databases and Information Systems Engineering Research’ in Bertino, E; Christodoulakis, S; Plexousakis, D; Christophies, V; Koubarakis, M; Bohm, K; Ferrari, E [Eds] Advances in Database Technology - EDBT 2004 Springer LNCS2992 pp3-16 ISBN 3-540-21200-0 March 2004


[OAI-PMH] Open Archives Initiative Protocol for Metadata Harvesting Homepage


[SIGLE]Open SIGLE Homepage http://opensigle.inist.fr/

/XML] W3C eXtended Markup Language Homepage http://www.w3.org/XML/