World Library and Information Congress:
71th IFLA General Conference and Council

"Libraries - A voyage of discovery"

August 14th - 18th 2005, Oslo, Norway

Conference Programme:
http://www.ifla.org/IV/ifla71/Programme.htm

Code Number:	**007-E**
Meeting:	**101 Science and Technology Libraries**

CRIS + Open Access = The Route to Research Knowledge on the GRID

Keith G Jeffery
Director IT,
CCLRC Rutherford Appleton Laboratory, Chilton, UK
E-mail: kgj@rl.ac.uk

Abstract

Linking CRISs (Current Research Information Systems) and OA (Open Access) Systems brings together systems for managing R&D with systems for providing open access to scholarly publishing – the major visible output of R&D – on the emerging European GRIDs infrastructure. The debate over OA is very active with 'green' (institutional repository self-archiving) and 'gold' (author / institution pays publishing) as competing but also complementary processes. The major publishers are experimenting with 'gold' services while 'green' institutional repositories are growing fast. GRIDs, especially through the NGG (Next Generation GRID) Reports [NGG] has emerged as a vision for a European IT 'surface' now being implemented progressively especially under the auspices of EC DG INFSO F2 to provide easy-to-use access to information and computation. CRISs provide both a context for evaluation of - and understanding the background to – scholarly publication. CRISs also provide a management framework for R&D in institutions from funding agencies through national laboratories to universities, as well as a mechanism for interoperating research and development information.

1 Introduction

We live in a world of competition and evaluation. The academic research process is not exempt: most countries evaluate the quality of universities and research institutes based on their measurable outputs – and make funding decisions based upon that analysis. One of the measurable outputs is research publications.

CRISs provide the information required to manage research and development, and also technology transfer and wealth creation. In the R&D world of today, CRISs are used widely in research funding organizations, research institutes and universities. However, they tend to record the context of the research – projects, researchers, organisations, equipment, funding – and not directly the outputs (patents, products, publications).

Publication of research results (or passing on of information or knowledge) has been a mainstay of the development of human technological culture for thousands of years, from cave-drawings of animals (intended perhaps to educate hunters or enthuse a hunting project) to the drawings and writings of Leonardo da Vinci (early technical reports). The availability of inexpensive printing provided an opportunity for explosive growth; some measure of quality was required. Learned societies criticised publications, often when in manuscript form and read to an audience, leading to the current peer-review process. Today a hyperlinked multimedia eprint with executable code and associated datasets may be reviewed by anyone adding an e-annotation.

The process of externalizing the concepts in the researcher's mind, of recording them (and associated experimental results or observations), preserves the result of the work beyond the lifetime of the researcher and also makes it replicable and distributable. Some claim that this 'preserved external memory' is the major distinguishing feature of humans.

It is to achieve this holistic view of the research – not only the outputs (of which the most obvious is publications) but also the context – that we wish to bring together CRISs and publications databases. More specifically, and taking advantage of the GRIDs technology, we wish to bring together CRISs and Open Access Repositories of publications so that e-Research is enabled. This then allows the end-user to have e-access to previous work, the research context, the experiment and its data / software. Where appropriate, the user may be able remotely to control / repeat experiments and produce new papers based on a solid platform of easily-available electronic material. This builds the corpus of research knowledge.

In parallel the research manager can analyse in context the research outputs and not just rely on, for example, simple counts of publications. Furthermore, the entrepreneur can utilise the CRIS data to find opportunities for wealth-creation.

The World Wide Web provides – via search engines - easy access to information, although with questionable relevance (accuracy) and recall (completeness). GRIDs technology – originating in metacomputing (linking together supercomputers to provide effectively a giant computer) – has now developed and adopted web services into OGSA (Open GRID Services Architecture). It is thus becoming seamless with the WWW and adds computation capability. The concept behind GRIDs in Europe (but not necessarily in North America) is that the end user makes a request and the overall GRIDs system interprets the request, proposes a 'deal' to the end user (which may involve money and/or trading of rights) and then, if accepted, executes the request using agents and brokers acting over heterogeneous sources and resources.

The scientific process can be treated as a workflow with recording of outputs at various stages, from initial ideas to project proposal to interim reports and final publications – along with the produced data, software and cross-references to other works.

This can all be synthesized in one pseudo-equation: CRIS + Open Access = The Route to Research Knowledge on the GRID. The rest of the paper explores Research Knowledge, CRISs, Open Access, GRIDs and finally synthesizes.

2 Research Knowledge

In a context of research, development and innovation, the IP (Intellectual Property) consists of products, patents and publications (in the widest sense – any stored representation of human intellect). While conventional research publications (white literature) provide much of the visible IP, the 'submerged part of the iceberg' is the organisation's grey literature. This commonly represents its 'know how' or knowledge base [Je99], [JeAsRe00). There are also legal considerations: many organizations protect their IPR with patents or pre-publication; copyright and database right are counterbalanced by Freedom of Information and Data Protection legislation. Innovation, Technology Transfer, Wealth Creation, Quality of Life are major objectives of R&D, and the reason why national governments, commercial organizations, charitable organizations and even individuals invest in it. Most of the technology upon which we depend today is the result of R&D years ago, and similarly the quality of life we enjoy is largely the result of R&D in topics such as medicine, education, environment. This is the IP.

Each organization needs to utilise its IP for business benefit (including the business of R&D) and for public relations / marketing purposes. This implies that an organisation needs to know the IP it owns, catalogue it, curate it and understand the business benefits from it – not least to inform investment decisions in future R&D to generate further IP. The knowledge base consists not only of the white literature but also the 'iceberg' of grey literature encapsulating the know-how of the organization in technical reports, instruction manuals, training materials etc. Furthermore, increasingly the IP rests in datasets (e.g. results of drug clinical tests), in databases (e.g. customer relationship information) and in software (which encapsulates the business processes of the organization). The whole may be subject to scrutiny through audit or freedom of information requests and thus there are great incentives for an organization to manage well these IP assets.

In the R&D world most public funding bodies now assess the output – IP - from an organisation that it has funded - e.g. a university or research institute - and they base future funding decisions – at least partly – on that information. Thus recording, curation and management of IP is critically important to research-based organisations. Similarly, for many modern businesses the quality of its IP determines success and future investment from shareholders. Thus we derive the requirement: to provide systems and an environment such that organisations can manage effectively their IP bringing together both the IP itself (e.g. grey literature) and the organisational business structures and objectives.

3 CRISs (Current Research Information Systems)

The historical development of CRISs (Current Research Information Systems) has emerged from the world of IR (Information Retrieval). The motivating reasons were that the recorded information was usually analogous to the {title/ author / date / keywords / address of source} form of library card catalogues.

Used stand-alone such systems were adequate until requirements for statistical analyses, integration with data in other DBMSs, flexible reporting (including integration with the client office environment) and handling of multimedia (or hypermedia) became important. These requirements emerged from the mid 1970s onwards. However, the emergence of the requirement to access information in multiple heterogeneous CRISs (and other DB systems) distributed geographically exposed harshly the inadequacies of these systems: they lacked any underlying theoretical model (to allow structural matching); they lacked common standards of data recording (to allow interworking) and they lacked common external interfaces (to allow integration with client office environments both for input/update and for retrieval/reporting). Worse, they lacked the metadata to allow such interworking to be provided 'on the fly' when required.

CERIF provides a comprehensive data model for R&D Information agreed by representatives of European countries (both European Union and associated States). The original CERIF1991 recommendation was used as the basis for the ERGO Pilot Project [ERGO]. The new CERIF data model was produced within the CERIF Revision Working Group [CERIF]. CERIF2000 provides a data model for projects, organizations, persons, funding, events, equipment, facilities, and all the relational linkages between them. This provides great flexibility allowing not only 1:n (hierarchic) relationships but also n:m (many-to-many or graph) relationships. Furthermore, the relationships are role-based and temporally defined, this providing a very rich data model. The model recorded the existence of - but assumed pre-existence of systems handling the detail for - patents, products and publications. CERIF2000 thus redefined CRIS in a formal way so providing a stable platform for CRIS builders, for inter-CRIS data exchange and for provision of metadata to describe CRIS contents succinctly.

The EC (European Commission) then handed care and development of CERIF to [EuroCRIS] since when later versions of CERIF have responded to users' needs in a structured way; now CERIF does include a detailed set of entities and attributes for publication information. The key proposals leading to this model are in [Je99], [JeAsRe00], [AsJe04].

4 OA (Open Access)

The WWW (World Wide Web) [W3C], has made e-publishing inexpensive and easy. This has led to an explosive growth of institutional (and subject-based e.g. [arXiv]) repositories. The Open Access Initiative [OAI] utilised the Dublin Core [DC] metadata standard and harvesting software (OAI-PMH) to link the repositories.

The two great challenges for the Web as outlined in [Be99] are the semantic web (to make the web understandable) and the web of trust (to make it secure). The semantic web is now being constructed, largely by (a) more formal data structures which are suitable for manipulation by first order logic, commonly involving the use of structured metadata and (b) use of domain ontologies to provide definitions of the meanings of terms and the logical inter-relationship of terms – as supportive associative metadata. The web of trust implies that material is secure from misuse and that the organization holding the information is trusted to utilise it in a way coincident with business ethics. This is achieved by associative restrictive metadata related to the original information.

Unfortunately, DC is not formalized, and so it is machine-readable but not machine-understandable. In many ways it is a step backwards from [MARC]. It lacks properties of the semantic web and web of trust. This limits severely its utility for automated processing, where formalised metadata is a key prerequisite. Metadata is data about data; a classification was proposed (originally in 1998, published in [Je00]) which ensures separation of metadata kinds and assists in correct logical processing. The application of the metadata classification to CRISs was demonstrated in [JeLoAs02]. A formalized version of DC was developed progressively in [Je99], [JeAsRe00], [AsJe04].

Open Access repositories imply that access is free at the point of demand. There are two major models: 'gold' open access is proposed mainly by publishers and requires the author (or author institution) to pay for publication. The alternative 'green' open access requires the author to deposit in a repository (institutional or subject-based) at or about the time of publication such that the material is freely available to online access. There are some problems over copyright with certain publishers, although the majority now allow open access with varying time-lags from their publication date. Of course 'green' open access also allows deposit of preprints and / or grey literature; this needs to be distinguished clearly from peer-reviewed publications. Meantime most publishers provide online access via subscription to their publications; the problem is that the researcher has to access multiple systems each with different interfaces (including login ID and password). For many the effort threshold is too high – they just use Google (or Google Scholar) and do not concern themselves with the relatively poor recall (completeness) or relevance (precision) of the search.

5 GRIDs

5.1 Background

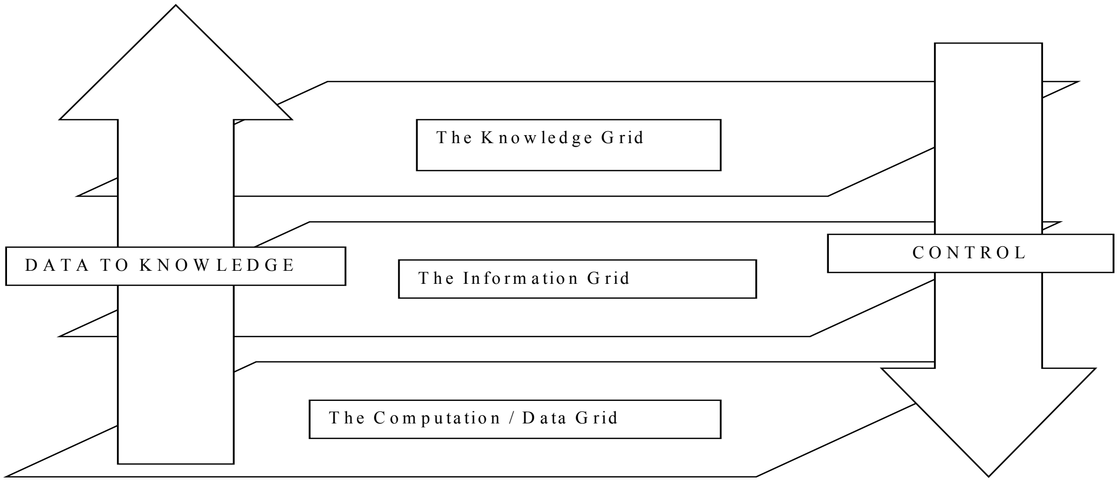

Figure 1: The 3-Layer GRIDs Architecture

In 1998-1999 the UK Research Council community of researchers was facing several IT-based problems. Their ambitions for scientific discovery included post-genomic understanding, climate change explanation, 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. The author was asked to propose an integrating IT architecture.

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. 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 which utilises knowledge discovery in database technology to generate knowledge and also allows for representation of knowledge through scholarly works, peer-reviewed (publications) and grey literature, the latter especially hyperlinked to information and data to sustain the assertions in the knowledge.

In parallel with the initial UK thinking on GRIDs, [FoKe98] published a collection of papers in a book generally known as 'The GRID Bible'. The essential idea is to connect together supercomputers to provide more power – the metacomputing technique. However, the major contribution lies in the systems and protocols for compute resource scheduling. The GRID corresponds to the lowest grid layer (computation / data layer) of the UK-proposed GRIDs architecture.

5.2 The GRIDs Architecture

The idea behind GRIDs is to provide an IT environment that interacts with the user to determine the requirement for service and then satisfies that requirement across a heterogeneous environment of data stores, processing power, special facilities for display and data collection systems thus making the IT environment appear homogeneous to the end-user.

The major components (
Figure 2) external to the GRIDs environment are: a) users: each being a human or another system; b) sources: data, information or software c) resources: such as computers, sensors, detectors, visualisation or VR (virtual reality) facilities. Each of these three major components is represented continuously and actively within the GRIDs environment by: 1) metadata: which describes the external component and which is changed with changes in circumstances through events 2) an agent: which acts on behalf of the external resource representing it within the GRIDs environment. Finally there is a component which acts as a 'go between' between the agents. These are brokers that, as software components, act much in the same way as human brokers by arranging agreements and deals between agents. From this it is clear that they key components are the metadata, the agents and the brokers.

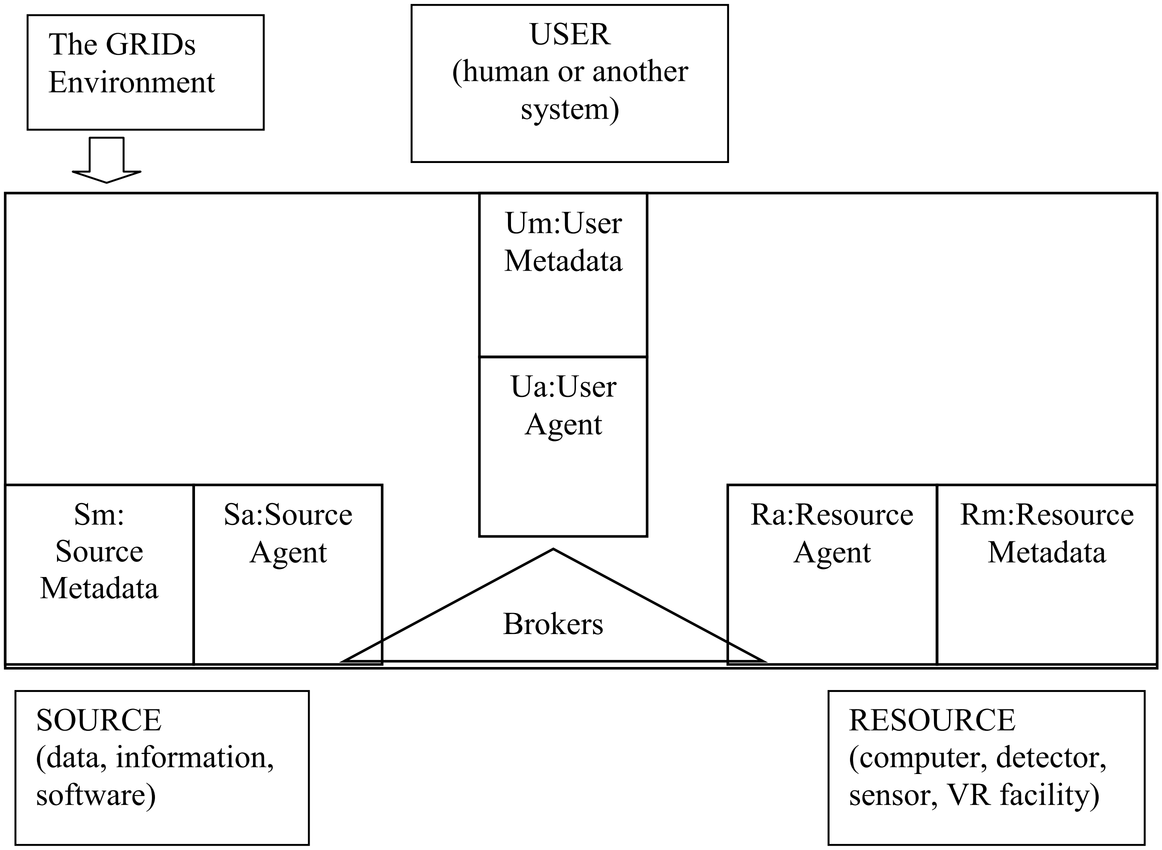

Figure 2: The GRIDs Components

5.3 Ambient Computing

The concept of ambient computing implies that the computing environment is always present and available in an even manner. The concept of pervasive computing implies that the computing environment is available everywhere and is 'into everything'. The concept of mobile computing implies that the end-user device may be connected even when on the move. In general usage of the term, ambient computing implies both pervasive and mobile computing.

A typical configuration might comprise: a) a headset with earphone(s) and microphone for audio communication, connected by bluetooth wireless local connection to b) a PDA (personal digital assistant) with small screen, numeric/text keyboard (like a telephone), GSM/GPRS (mobile phone) connections for voice and data, wireless LAN connectivity and ports for connecting sensor devices (to measure anything close to the end-user) in turn connected by bluetooth to c) an optional notebook computer carried in a backpack (but taken out for use in a suitable environment) with conventional screen, keyboard, large hard disk and connectivity through GSM/GPRS, wireless LAN, cable LAN and dial-up telephone.

The end-user would perhaps use only (a) and (b) (or maybe (b) alone using the built in speaker and microphone) in a social or professional context as mobile phone and 'filofax', and as entertainment centre, with or without connectivity to ‘home base’ servers and IT environment. For more traditional working requiring keyboard and screen the notebook computer would be used, probably without the PDA. The two might be used together with data collection validation / calibration software on the notebook computer and sensors attached to the PDA.

Such a configuration is clearly useful for a ‘road warrior’ (travelling salesman), for emergency services such as firefighters or paramedics, for businessmen, for production industry managers, for the distribution / logistics industry (warehousing, transport, delivery), for scientists in the field…. and also for leisure activities such as mountain walking, visiting an art gallery, locating a restaurant or visiting an archaeological site. The concept is access to a GRIDs environment anyhow, anytime, anywhere. Ambient computing brings the power of GRIDs to the hand. Linked with CRISs and Open Access repositories it provides the researcher, research manager or entrepreneur with easy access to the knowledge required.

6. Synthesis

Thus, The GRIDs concept – initially metacomputing i.e. linking supercomputers [FoKe98] - has been extended (initial internal papers in 1999 and published in [Je01]) to a full-blown distributed computing environment including, as GRID services, the W3C concept of web services together with concepts of the semantic web and web of trust. Through ambient computing the environment is available anywhere and through any device of choice [Je04].

This environment provides the platform for the ultimate blurring of white and grey literature, from refereed publications through annotated preprints to technical reports and manuals with formalized metadata allowing automated processing in Open Access repositories. This is cross-linked to datasets and appropriate software and – using CRIS technology [EuroCRIS] - to persons, organizations, projects, patents, publications, events, facilities and equipment. The environment is completed with associated computation power, special output facilities (e.g. VR (Virtual Reality) and dynamic control of detectors and instrumentation collecting data. It supports the complete research process as a workflow.

At CCLRC we have built:

(a) an institutional open access repository with over 20,000 entries in formalized DC to allow automated processing and interoperation; the data model also utilises the concepts from the [IFLA FRBR] model;

(b) a corporate data repository using the CERIF datamodel (extended to handle internal business process management as well as CRIS requirements);

(c) a GRIDs environment with ambient computing access allowing easy interoperation and easy user access to information and computation facilities;

We are now integrating them in parallel with re-engineering the organisation’s business processes to provide the e-Research environment for the future. This will provide an effective and efficient basis for the management of R&D.

Acknowledgements

Most of the work on CRISs has been done in conjunction with colleagues in euroCRIS, whose contributions I acknowledge. The major work on CERIF and on linking publications information to CRISs has been done jointly with Anne Asserson, University of Bergen. Most of the work on Open Access repositories has been done with my own team at CCLRC –

particularly Matthew Mascord, Catherine Jones, Brian Matthews and Heather Weaver - who provide a continual source of inspiration and ideas. The work on GRIDs has benefited additionally from discussions with the UK e-Science community and wider discussions in a European context, mainly through [ERCIM] and the [NGG] Group. The synthesis and this paper are my responsibility.

References

[ArXiv] www.arxiv.org

[AsJeLo02] Asserson,A; Jeffery,K.G; Lopatenko,A: '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 2002 (available under www.eurocris.org)

[AsJe04] Asserson, A; Jeffery, K.G.; 'Research Output Publications and CRIS' in A Nase, G van Grootel (Eds) Proceedings CRIS2004 Conference, Leuven University Press ISBN 90 5867 3839 May 2004 pp 29-40 (available under www.eurocris.org)

[Be99] Berners-Lee,T; 'Weaving the Web' 256 pp Harper, San Francisco September 1999 ISBN 0062515861

[CERIF] http://www.cordis.lu/cerif/

[DC] http://purl.oclc.org/metadata/dublin_core/

[ERCIM] http://www.ercim.org

[ERGO] http://www.cordis.lu/ergo/

[EuroCRIS] http://ww.eurocris.org

[FoKe98) Foster I and Kesselman C (Eds). The Grid: Blueprint for a New Computing Infrastructure. Morgan-Kauffman 1998

[IFLA] http://www.ifla.org/

[IFLA FRBR] http://www.ifla.org/VII/s13/frbr/frbr.pdf

(Je99) Jeffery, K G: 'An Architecture for Grey Literature in a R&D Context' Proceedings GL'99 (Grey Literature) Conference Washington DC October 1999
http://www.konbib.nl/greynet/frame4.htm

[JeAsRe00] Jeffery K.; Asserson A.; Revheim J; (2000) CRIS, Grey Literature and the Knowledge Society, Proceedings CRIS-2000, Helsinki
ftp://ftp.cordis.lu/pub/cris2000/docs/jeffery_fulltext.pdf

[Je00] Jeffery, K.G., 2000, 'Metadata': in Brinkkemper,J; Lindencrona,E; Solvberg,A: 'Information Systems Engineering' Springer Verlag, London 2000. ISBN 1-85233-317-0.

[Je01] Jeffery,K.G.;'GRIDs: Next Generation Technologies for the Internet' Invited Keynote Presentation (Abstract in Proceedings) Eds Wang,Y; Patel,S; Johnston,R.H; OOIS2001 Conference, Calgary August 2001, page 1, Springer, ISBN 1-85233-546-7

[JeLoAs02] Jeffery,K.G; Lopatenko,A; Asserson, A.: '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 (available at www.eurocris.org)

[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

[Je04a] Jeffery, K.G.; 'The New Technologies: can CRISs Benefit' in A Nase, G van Grootel (Eds) Proceedings CRIS2004 Conference, Leuven University Press ISBN 90 5867 3839 May 2004 pp 77-88 (available at www.eurocris.org)

[MARC] http://minos.bl.uk/services/bsds/nbs/marc/commarcm.html

[NGG] www.cordis.lu/ist/grids

[OAI] www.openarchives.org

[W3C] www.w3.org

[W3Cmetadata] http://www.w3.org/Metadata/