CRIS AND THE GRIDS ARCHITECTURE

K 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

The end-user demands low effort threshold access to systems providing e-information, e-business, and e-entertainment. Innovators and entrepreneurs require also equally low-energy access to heterogeneous information homogenised to a form and language familiar to them. On top of that, decision-makers, whether in a control room or government strategic planning, demand equally easy access to information that is statistically or inductively enhanced to knowledge and access to modelling or simulation systems to allow 'what if?' requests. Researchers and technical workers have an additional requirement for rapid integration of information with statistical, induction, modelling, and simulation systems to generate and verify hypotheses so generating data and information, to be used by others, which in turn advances knowledge. Access is required, and can now be provided, anytime, anyhow, anywhere through ambient computing technology. A new paradigm, GRIDs, provides the architectural framework.

1 INTRODUCTION

IT, from microstructures (chips) through telecommunications to knowledge-based information handling, is developing very fast. Effectively unlimited digital storage, computer capacity, and bandwidth conspire to make the constraints on business disappear. Literally, we are limited only by our imagination, and with advanced internet n-dimensional games even this is being stretched.

Against this landscape, new technologies are emerging. They offer effectively unlimited scope for the development of business. Legislation, ethics, and human understanding of the possibilities to use the technology struggle to keep up with the pace of technology change.

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 visualization, all supported by easy-to-use tools controlled through an intuitive user interface. The author was asked to propose an integrating IT architecture.

The idea depended on the idea of bringing together in one easy-to-use environment observational data, experimental data, and the outputs from simulation or modelling into one environment for the researcher to analyse and utilize, together with facilities for the usual researcher tasks, such as research proposals, project management, publishing, etc. The results (including the raw data) would then be widely available 'open access,' subject only to rights such as investigator priority. This architectural idea is equally applicable to the business environment, but here we are concerned with the research environment. Specifically, we are concerned with the relationship of the GRIDs technology to CRIS (Current Research Information Systems) particularly those based on CERIF (Common European Research Information Format)

2 GRIDS

2.1 Architecture

The architecture proposed (Jeffery, 1999a) 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 and peer-reviewed (publications) and grey literature, the latter especially hyperlinked to information and data to sustain the assertions in the knowledge (Jeffery, 1999a), (Jeffery et al., 2000).

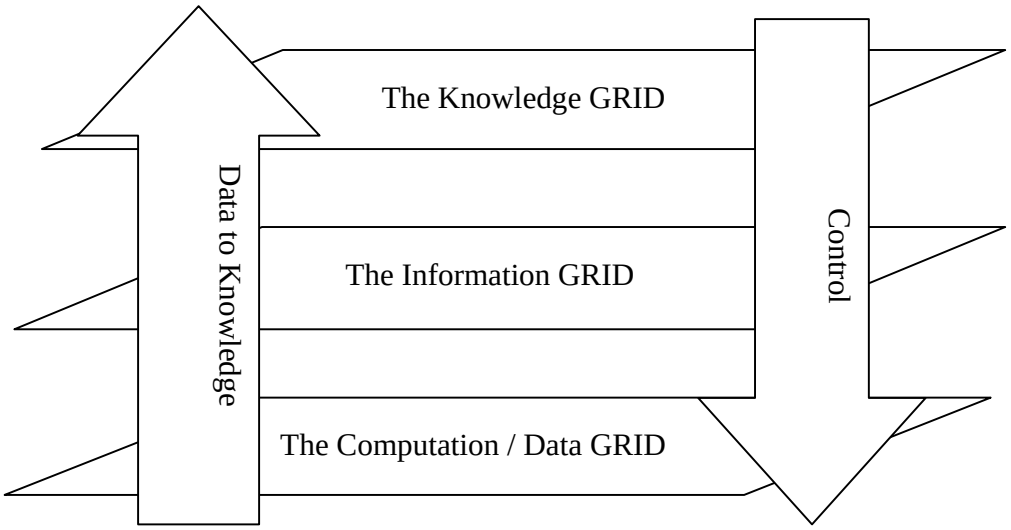

Figure 1. The 3-Layer GRIDs Architecture

In parallel with the initial UK thinking on GRIDs, Foster and Kesselman (1998) 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 computer resource scheduling. The GRID corresponds to the lowest grid layer (computation / data layer) of the UK-proposed GRIDs architecture.

2.2 Components

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, visualization, or VR (virtual reality) facilities. Each of these three major components is represented continuously and actively within the GRIDs environment by: 1) metadata: that describes the external component and which is changed with

changes in circumstances through events and 2) an agent: that acts on behalf of the external resource representing it within the GRIDs environment. Finally there is a component that 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.

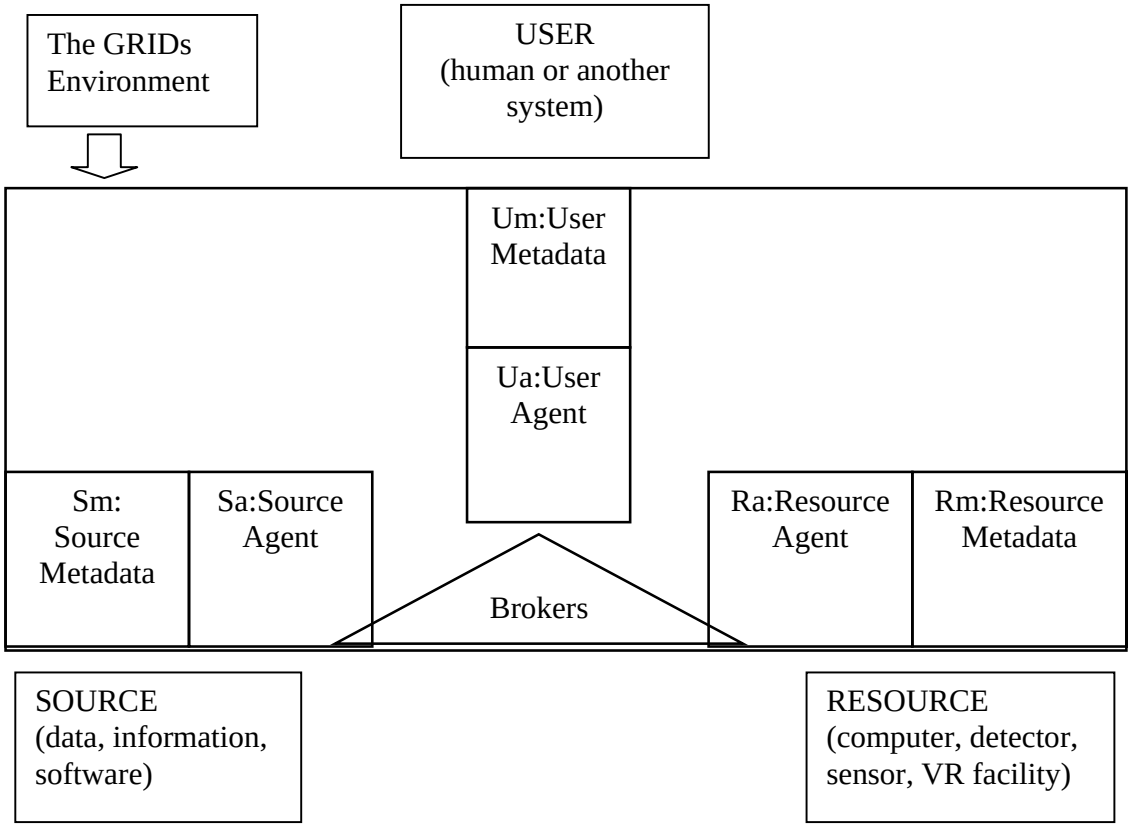

Figure 2. The GRIDs Components

2.3 Metadata

Metadata is data about data (Jeffery, 2000). An example might be a tag attached to a museum specimen. The metadata on the tag tells the end-user (human examining the specimen) data about the article itself, such as the location and date of discovery, the identification, and classificatory information. The metadata tag may be attached directly to the specimen, or it may appear in a catalogue of the museum collection (or, more usually, both). The metadata may be used to make a selection of potentially interesting specimens before the actual specimens are inspected, thus improving convenience. Today this concept is widely-used. Much e-commerce is based on B2C (Business to Customer) transactions based on an online catalogue (metadata) of goods offered. One well-known example is www.amazon.com .

It is increasingly accepted that there are several kinds of metadata. The classification proposed initially in 1998 (Jeffery, 2000) (Figure 3) is gaining wide acceptance and is detailed below.

Schema metadata constrain the associated data. One problem with existing schema metadata (e.g., schemas for relational DBMSs) is that they lack certain intentional information that is required (Jeffery et al., 1994). Systems for information retrieval based on, e.g., the SGML (Standard Generalised Markup Language) DTD (Document Type Definition) experience similar problems. It is noticeable that many ad hoc systems for data exchange between systems send with the data instances a schema that is richer than that in conventional DBMS, to assist the software (and people) handling the exchange to utilise the exchanged data to best advantage.

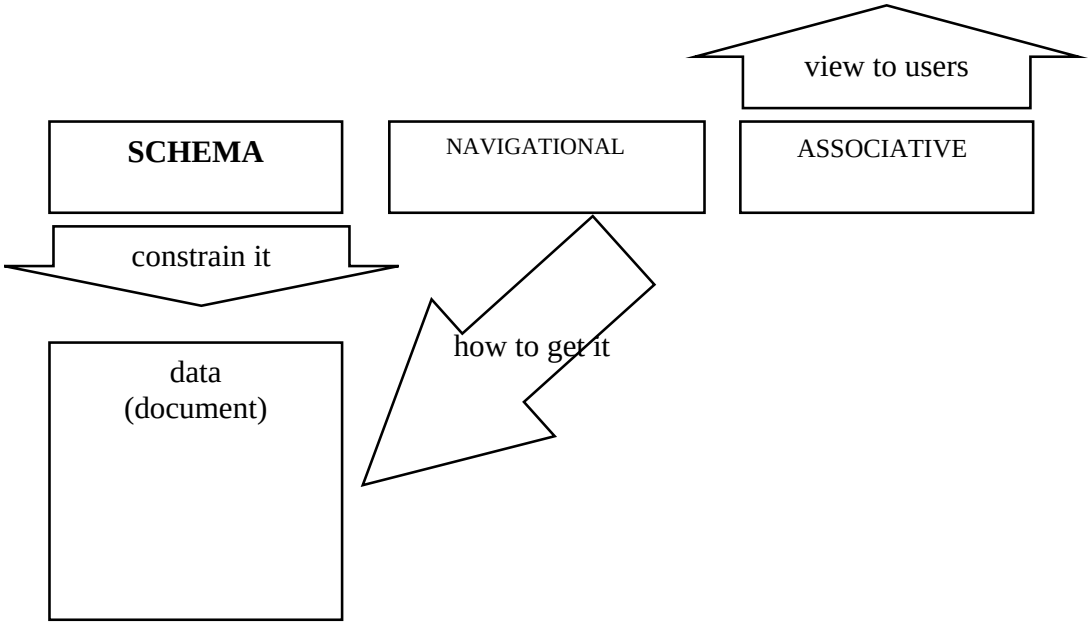

Figure 3. Metadata Classification

Navigational metadata provides the pathway or routing to the data described by the schema metadata or associative metadata. In the RDF model, it is a URL (universal resource locator), or more accurately, a URI (Universal Resource Identifier). With increasing use of databases to store resources, the most common navigational metadata now is a URL with associated query parameters embedded in the string to be used by CGI (Common Gateway Interface) software or proprietary software for a particular DBMS product or DBMS-Webserver software pairing.

Associative metadata may be classified as follows: 1) descriptive: provides additional information about the object to assist in understanding and using it; 2) restrictive: provides additional information about the object to restrict access to authorised users and is related to security, privacy, access rights, copyright, and IPR (Intellectual Property Rights); and 3) supportive: a separate and general information resource that can be cross-linked to an individual object to provide additional information, e.g., translation to a different language, super- or sub-terms to improve a query, the kind of support provided by a thesaurus or domain ontology.

Most examples of metadata in use today include some components of most of these kinds but neither structured formally nor specified formally so that the metadata tends to be of limited use for automated operations, particularly interoperation, thus requiring additional human interpretation.

The mapping of this metadata scheme to CRISs was described in (Jeffery, Lopatenko, & Asserson, 2002).

2.4 Agents

Agents operate continuously and autonomously and act on behalf of the external component they represent. An agent's actions are controlled to a large extent by the associated metadata, which should include either instructions,

or constraints, such that the agent can act directly or deduce what action is to be taken. Each agent is waiting to be 'woken up' by some kind of event; on receipt of a message the agent interprets the message and, using the metadata as parametric control, executes the appropriate action, either communicating with the external component (user, source, or resource) or with brokers as a conduit to other agents representing other external components.

2.5 Brokers

Brokers act as 'go betweens' between agents. Their task is to accept messages from an agent requesting some external component (source, resource, or user), identify an external component that can satisfy the request by its agent working with its associated metadata, and either put the two agents in direct contact or continue to act as an intermediary, possibly invoking other brokers (and possibly agents) to handle, for example, measurement unit conversion or textual word translation.

2.6 Interaction

Now let us consider how the components interact. An agent representing a user may request a broker to find an agent representing another external component, such as a source or a resource. The broker will usually consult a directory service (itself controlled by an agent) to locate potential agents representing suitable sources or resources. The information will be returned to the requesting (user) agent, probably with recommendations as to order of preference based on criteria concerning the offered services. The user agent matches these against preferences expressed in the metadata associated with the user and makes a choice. The user agent then makes the appropriate recommendation to the end-user, who in turn decides to 'accept the deal' or not. As well as this 'pull' technology, the user agent can set up a monitoring activity such that any change in the world of information of interest to the end-user causes an alert or regular update bulletin to be sent to the end-user.

3 AMBIENT COMPUTING

3.1 Concept

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.

3.2 Configuration

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, or dial-up telephone.

3.3 Use

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.

4 EVOLUTION OF CRIS

The technology is only one aspect: for maximum benefit CRISs also need to evolve. The new technologies overcome technical and economic limitations, which have previously restricted CRISs to limited data on projects or expertise. The end-user will expect to discover knowledge from information extended both in depth (more detailed data relating to the entity of interest) and breadth (more instances of data thus ensuring a better representation of the real world). In the case of greater depth, the end-user should be able to obtain not only information on projects, persons, and organizations and their patents, products and publications (i.e., the scope of CERIF) (CERIF) but also the actual publications online (Jeffery & Asserson 2004; 2005) with references to the data upon which the work is based and any associated software, instrumentation, methods, and techniques (Jeffery & Asserson, 2006a; 2006b). In the case of breadth, better and easier data recording, based largely on e-forms, re-use, and instrumentation will provide for the end-user a broader information landscape. Technology will lower the effort threshold for data collection and also increase the interconnection of information sources. There are dangers that these improvements may be haphazard and unstructured, for example, by end-user harvesting of CRIS information directly without any mediation or conversion.

There is an urgent need for CRISs to evolve such that they provide for the end-user all the information required in a form suitable for further knowledge-processing. This requires existing CRISs to expand considerably the scope of their schemas both in breadth and depth. For some of these extensions, international standards already exist (e.g., in certain areas of scientific information exchange and in bibliography / bibliometrics), and these should be adopted as extensions linked to CERIF. In other areas, such standards do not exist or multiple standards exist (e.g., publication formats, grey literature, documentation of methods and process (Jeffery, 1999b), and documentation of patents). In this area more work needs to be done to achieve harmony. In both cases, it is likely that the extended information will be available in other information systems so that the requirement is that the CERIF CRIS should provide a seamless linkage such that the end-user does not realize the information is not coming from the original CRIS (Jeffery & Asserson 2006a; 2006b). Such linkage is already usable between CERIF and non-CERIF CRISs (Jeffery, 2005) and so, by knowledge-based domain extension technology, the proposed linkages should not be a problem.

5 POTENTIAL USE FOR CRIS

Following our assertion, we now discuss how GRIDs and Ambient Computing offer the potential for development of business associated with CRIS. The key benefits from these technologies used together are: (a) anywhere-, anytime-, anyhow-access through pull or push technology; (b) knowledge-assisted user profiling to improve relevance and recall of information and to meet user preferences; (c) provision of computation as well as information so allowing integrated processing such as statistics, data mining, or visualization, for e.g., decision support; (d) homogeneous view over heterogeneous information; and (e) handling of rights and legalities. The really new concept is that the system interacts with the end-user to determine the requirement and then in real time puts together the components needed to satisfy the request, delivering the result conveniently.

The new technology makes it much easier to present the material usually stored in a CRIS, or accessed via a CRIS, in an attractive way, tuned to the end-user need. Database-generated web pages (XML) describing the usual CERIF entities can be available, translated as necessary, hyperlinked together, presented through XSLT (thus allowing transformations such as presentation on a widescreen device, a laptop or PDA or even as audio through a mobile phone) and furthermore allowing the end-user to obtain statistical summaries by various key parameters such as subject area, geographical region, and temporal region. The use of the user profile improves dramatically relevance, recall, presentation preferences, and effort threshold.

People working in the media want newsworthy, interesting, brief and well-presented material. The new technologies, especially with advanced visualisation, augmented with virtual reality and modeling / simulation capabilities can provide this. Of course, the data provided must represent accurately the real world; despite prejudices to the contrary, in fact, media people do rely on trustworthy sources. For media people, the use of push technology for alerts on

changes in CRIS information, especially in currently hot topics such as post-genomics, potential asteroid collisions with earth, climate change, or new technology.

Entrepreneurs and innovators demand current information on ideas, discoveries, and technologies that are exploitable for wealth creation or improving the quality of life. Moreover, the entrepreneur requires detailed understanding of the information from perspectives of the technical reliability of the features, the market potential of the benefits, and associated rights and legal issues. The new technologies are almost an ideal match for this requirement since they permit the deep searching necessary, provide analysis systems and modelling facilities, and associated visualization. They also assist in searching (via restrictive metadata) for rights and legalistic aspects.

The decision maker usually wishes to have a knowledge-based assist system which acts as an expert advisor, not executing the recommended course of action but advising the end-user. Such a system is based on knowledge-based technology for weighting and optimizing possible strategies. The strategies themselves are based on analysis of information, together with modelling or simulation facilities to allow for 'what if?' options to be considered. Such a system also requires access to the maximum amount of relevant information and access to computation facilities to calculate options over the data. Of course, inference processing is also required for the knowledge processing aspects, controlled by metadata and agents representing all the components, including the end-user. Visualisation, augmented reality, and virtual reality can all assist the decision-maker in understanding the current state of the world of interest and the proposed courses of action and their effects. Finally, decision-makers sometimes make less complicated decisions when traveling, so presentation of the options and recommendations through a PDA is also important; the decision-maker can demand more in-depth information if required.

The new technologies are, to some extent, designed for the researcher. The lowered effort threshold to integrate data into information, to analyse, model and simulate, to display using visualization, augmented reality or virtual reality, and the ability to attempt repeated hypothesis testing in a short time all provide an amazing facility for the researcher. The new technology also permits control over the internet of scientific instruments and detectors on satellites, major facilities, and throughout the natural environment. The technology increases researcher productivity dramatically and causes the researcher to adopt a new way of doing research, indeed a new business opportunity created by the technology.

6 THE R&D TO ADVANCE GRIDs

There is much ongoing R&D in GRIDs. The major challenges concern representativity (of the real world with associated concepts of integrity), performance (including resource scheduling, business continuity, and mobile code), usability (novel user interfaces, high-level declarative languages), and security and trust including privacy. More recently some advanced research has been investigating taking the SOA (Service-Oriented Architecture) ideas and having metadata around the services to permit discovery, utilization, and management (performance, service-level agreements, trust, and security) using intelligent, knowledge-based techniques. This research moves well beyond traditional web services management through WSDL (Web Service Description Language), BPEL (Business process Execution Language), etc.

These advanced services are described as SOKU (Service-Oriented Knowledge Utility), and the idea originated in the Next Generation Expert Group of the EC in 2006. However, with the emergence of ambient computing and 'the Internet of Things' there are potentially billions of nodes on the network from supercomputers (or clusters as Clouds) to tiny intelligent detectors associated with research experimentation or observation. Such an environment challenges the very basis of computer science and information systems engineering. To indicate the areas of problems, consider the concept of state across millions of nodes and the problems of update transactions on distributed heterogeneous databases with thousands of nodes, each of which is simultaneously being updated locally in real-time with streams of data from detector arrays.

These advances when formalised will assist greatly in the research information environment by making software development more reliable, faster, and cheaper and subsequent maintenance easier while providing greater performance and improved trust, security, privacy, service-level guarantees, and business continuity.

7 CONCLUSIONS AND RECOMMENDATIONS

CRIS designers and implementers should aim to adopt in an evolutionary manner the new technologies in order to provide for their end-users the benefits, from low effort threshold to new business opportunities. The key steps are: a) the metadata description of the users, sources, and resources and b) the extension of CERIF metadata to describe separate but linked sources of detailed scientific, bibliographic, and other relevant data in order to provide the end-user with a homogeneous view over heterogeneous in-depth data. The euroCRIS community should be involved actively in the definitions of these metadata standards to assure the interests of the CRIS community

8 REFERENCES

Asserson, A, Jeffery, K.G., & Lopatenko, A. (2002) 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.

Asserson, A. & Jeffery, K.G. (2004) Research Output Publications and CRIS. In Nase, A. & van Grootel, G. (Eds.) *Proceedings CRIS2004 Conference*, Leuven University Press ISBN 90 5867 3839, pp 29-40.

Asserson, A. & Jeffery, K.G. (2005) Research Output Publications and CRIS. *The Grey Journal* 1 (1). TextRelease/Greynet ISSN 1574-1796, pp 5-8.

CERIF data model documentation: Retrieved from the World Wide Web 20 April 2010 http://www.eurocris.org/cerif/introduction/

DC Retrieved from the World Wide Web, April 13, 2010: http://dublincore.org/

Foster, I. & Kesselman, C. (Eds) (1998) *The Grid: Blueprint for a New Computing Infrastructure*. Morgan-Kauffman.

Jeffery, K.G. (1999a) Knowledge, Information and Data. Paper submitted to Director General of Research Councils. Retrieved from the WWW, April 13, 2010: http://www.semanticgrid.org/docs/KnowledgeInformationData/KnowledgeInformationData.html

Jeffery, K G. (1999b) An Architecture for Grey Literature in a R&D Context. *Proceedings GL'99 (Grey Literature) Conference* Washington DC. Retrieved from the WWW April 20 2010 http://www.emeraldinsight.com/Insight/viewContentItem.do?contentType=Article&contentId=861814

Jeffery, K G. (2000) Metadata. In Brinkkemper, J, Lindencrona, E., & Solvberg, A (Eds.) *Information Systems Engineering.* Springer Verlag, London. ISBN 1-85233-317-0.

Jeffery, K.G. (2004) The New Technologies: can CRISs Benefit. In Nase, A. & van Grootel, G. (Eds.) *Proceedings CRIS2004 Conference*, Leuven University Press ISBN 90 5867 3839, pp 77-88.

Jeffery, K. G. (2005) CRISs, Architectures and CERIF. *CCLRC-RAL Technical Report* RAL-TR-2005-003.

Jeffery, K. G. & Asserson, A. (2006) CRIS Central Relating Information System. In Asserson, A. & Simons, E. (Eds.) Enabling Interaction and Quality: Beyond the Hanseatic League. *Proceedings 8th International Conference on Current Research Information Systems CRIS2006 Conference*, Bergen, Leuven University Press ISBN 978 90 5867 536 1, pp109-120.

Jeffery, K. G. & Asserson, A. (2006b) Supporting the Research Process with a CRIS In Asserson, A. & Simons, E. (Eds.) Enabling Interaction and Quality: Beyond the Hanseatic League. *Proceedings 8th International Conference on Current Research Information Systems CRIS2006 Conference*, Bergen, Leuven University Press ISBN 978 90 5867 536 1, pp 121-130.

Jeffery, K.G, Lopatenko, A, & Asserson, A. (2002) 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.

Jeffery, K. G., Hutchinson, E. K., Kalmus, J. R., Wilson, M. D., Behrendt, W., & Macnee, C. A. (1994) A Model for Heterogeneous Distributed Databases. *Proceedings BNCOD12,* LNCS 826 Springer-Verlag, pp 221-234.

Jeffery, K G., Asserson, A., Revheim, J., & Konupek, H. (2000) CRIS, Grey Literature and the Knowledge Society. *Proceedings CRIS2000 Conference*, Helsinki, Finland. Retrieved from the World Wide Web 20, April 2010: ftp://ftp.cordis.europa.eu/pub/cris2000/docs/jeffery_fulltext.pdf

Lopatenko, A., Asserson, A., & Jeffery, K.G CERIF: Information Retrieval of Research Information in a Distributed Heterogeneous Environment. In Adamczak, W. & Nase, A. (Eds.) *Proceedings CRIS2002 6th International Conference on Current Research Information Systems.* Kassel University Press ISBN 3-0331146-844, pp 59-68.