

The New Technologies: can CRISs Benefit?

KEITH G. JEFFERY
CLRC

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 homogenized 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 modeling 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. Where are the CRISs?

1 Business and Technology

1.1 The Assertion

Conventional wisdom concerning the development of IT (information technology) is as follows: a) define the mission and strategic objectives of the organization; b) define the means by which they are to be achieved using business management techniques; c) provide the IT to support b). However, present-day business success depends much more on an interplay between business management techniques and IT. For example, amazon.com is a business that exists only because of the technology; the originators saw how they could harness the technology to create a successful business organisation. Many present-day business organisations could not operate as effectively or efficiently without having changed their business to adopt the technology: examples include internet banking and online travel reservations. Others still follow the conventional wisdom but nonetheless have efficiency gains with IT over using manual business processes.

1.2 Implication for CRISs

The assertion is that we should stop trying to define the business requirements and then build supportive CRISs but instead look to IT for opportunities to create new

business with CRISs. CRISs will evolve to provide more in-depth information and broader (more complete) information.

2 The Technologies

2.1 Introduction

IT, from microstructures (chips) through telecommunications to knowledge-based information handling is developing very fast. Effectively unlimited digital storage, compute 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.

2.2 GRIDs

2.2.1 Background

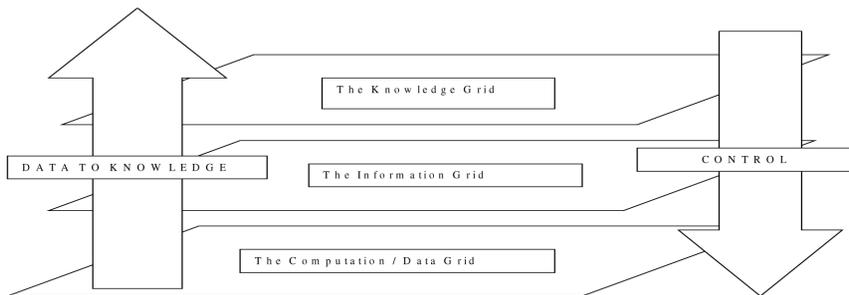


Figure 1: The 3-Layer GRIDs Architecture

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 (Jeffery 1999), (Jeffery, Asserson & Revheim 2000).

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 compute resource scheduling. The GRID corresponds to the lowest grid layer (computation / data layer) of the UK-proposed GRIDs architecture.

2.2.2 The GRIDs Architecture

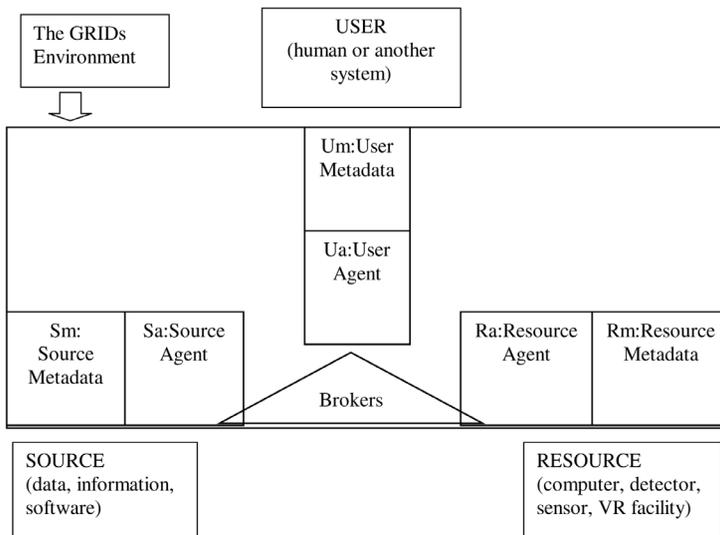


Figure 2: The GRIDs Components

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 (Fig.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 which, 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.

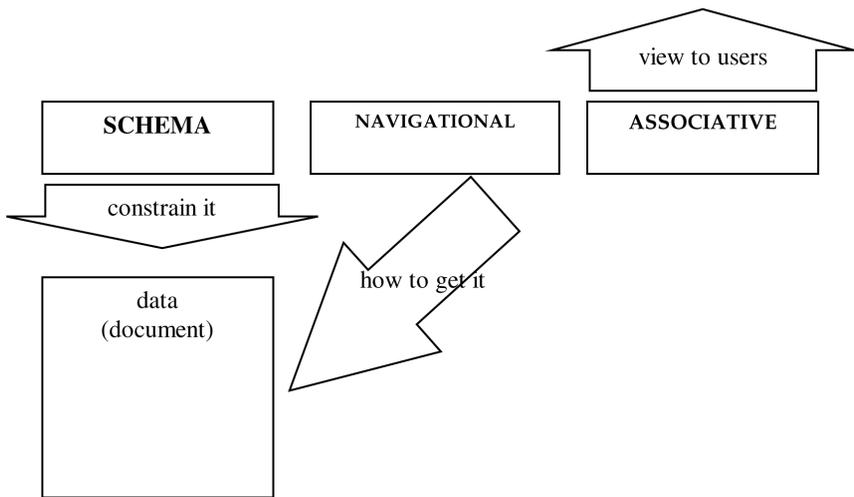


Figure 3: Metadata Classification

2.2.3 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 (Jeffery 1998) (Fig. 3) is gaining wide acceptance and is detailed below.

Schema metadata constrains the associated data. One problem with existing schema metadata (e.g. schemas for relational DBMS) is that they lack certain intensional information that is required (Jeffery and others, 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.

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); 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).

2.2.4 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.

2.2.5 Brokers

Brokers act as ‘go betweens’ between agents. Their task is to accept messages from an agent which request 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.

2.2.6 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.

2.3 Ambient Computing

2.3.1 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.

2.3.2 Configuration

A typical configuration might comprise: a) a headset with earphone(s) and microp-

hone 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.

2.3.3 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.

3 Evolution of CRISs

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) but also the actual publications online with references to the data upon which the work is based and any associated software, instrumentation, methods and techniques. 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, documentation of patents). In this area more work needs to be done to harmonise. 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. Such linkage is already usable between CERIF and non-CERIF CRISs and so – by knowledge-based domain extension technology – the proposed linkages should not be a problem.

4 Potential Use for CRIS

4.1 Introduction

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; (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.

4.2 General Public

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 webpages (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, temporal region. The use of the user profile improves dramatically relevance, recall, presentation preferences and effort threshold.

4.3 Media

People working in the media want newsworthy, interesting, brief and well-presented material. The new technologies, especially with advanced visualisation, augmented and 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.

4.4 Entrepreneur

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.

4.5 Decision-Maker

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 modeling 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 trave-

ling – so presentation of the options and recommendations through a PDA is also important; the decision-maker can demand more in-depth information if required.

4.6 Researcher

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 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.

5 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 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.

6 References

- Foster, I and Kesselman, C (Eds). 1998, *The Grid: Blueprint for a New Computing Infrastructure*. Morgan-Kauffman
- Jeffery, K.G., Gill, E.M., 1976 *The Design Philosophy of the G-EXEC System*. *Computers and Geosciences* 2(1976) 345-346.
- Jeffery, K; Hutchinson, E; Kalmus, J; Wilson, M; Behrendt, W; Macnee, C (1994) *A Model for Heterogeneous Distributed Databases* Proceedings BNCOD12 July 1994; LNCS 826 pp 221-234 Springer-Verlag 1994
- Jeffery, K (1998) *The Future of CRIS* Invited Paper CRIS98 Conference, March 1998, Luxembourg. <http://www.cordis.lu/cybercafe/src/jeffery.htm>
- Jeffery, K (1999) *An Architecture for Grey Literature in a R&D Context* Proceedings of Grey Literature Conference, 1999, Washington, US

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

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

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

van Woensel, L: (1988) CERIF Manual October 1988

W3C www.w3.org

W3Cmetadata <http://www.w3.org/Metadata/>

(XML) see (W3C)

7 Contact Information

e-mail: keith.g.jeffery@rl.ac.uk