GRIDS AND AMBIENT COMPUTING FOR GEOSCIENCE

Keith G Jeffery
Director, IT; CLRC Rutherford Appleton Laboratory, UK
k.g.jeffery@rl.ac.uk

Introduction

In the period 1968-1971 an IT system for geology (EGSTAR) was developed by the author at Exeter University. Its main characteristics were a consistent formal user command language to invoke processes, and the interconnection of processes through self-describing files. More formally, and with hindsight, it was an implementation of the relational algebra with additions towards computational completeness. The system was much refined and extended at the Institute of Geological Sciences (now British Geological Survey) with a team including Elizabeth Gill, Steve Henley and John Cubitt and was used extensively from 1972 under the name G-EXEC (Jeffery and Gill 1976). It provided a batch processing environment for data input, validation, storage, retrieval, transformation, analysis, visualisation, modelling and mining. The user interaction was later made interactive: a front-end editing system (G-FILE) was added, to be followed by an intelligent user interface guided by rules and learning (MARVIN).

Thirty years on, we live in a complex and fast-changing world. There is a tidal wave of data from sensors and video cameras. Users require information and knowledge to be extracted from these vast volumes of data in such a way as to assist human decision-making: it must be correct, in appropriate form, at the right place in a timely fashion. The overall requirement is to collect data which represents accurately the real world, and then to process it to information (structured data in context) from which can be extracted knowledge (justified commonly held belief). This, together with a cooperative working environment, provides the capability for decision support and problem solving. It also provides an environment for enjoyment of culture and education. The next generation GRIDs technology combined with Ambient Computing meets the need.

Since proposing formally in 1999 the idea described here, the author and his team have been actively researching the area. It combines well with our work managing the UK and Ireland Office of W3C (the World Wide Web Consortium) and associated R&D work on standards. The idea is not at all dissimilar to the original EGSTAR / G-EXEC concept of 30 years ago.

The Problem

Humans have to search for information from many sources and find it stored in different character sets, languages, data formats, in different measurement units, at different precision, with different accuracies. Worse, having found appropriate sources they have to find sufficient computing power to process the data, to integrate it together to a useful form for the problem at hand and suitable visualisation capabilities to provide a human-understandable view of the information. Suitable expert decision support systems and data mining tools for the creation of knowledge from information, and communications environments for group decision-making, are also hard to find and use. Behind all this is the need for accurate (represents accurately the world of interest) and precise (measured correctly) data; this requires appropriate data collection devices, supporting software and validation systems. The new paradigms of GRIDs and Ambient Computing are an attempt to overcome these problems.

The paper is organised as follows: in succeeding sections the GRIDs concept is described. The architecture is detailed. Ambient Computing is described. In the final Section the research issues are explored.

GRIDs

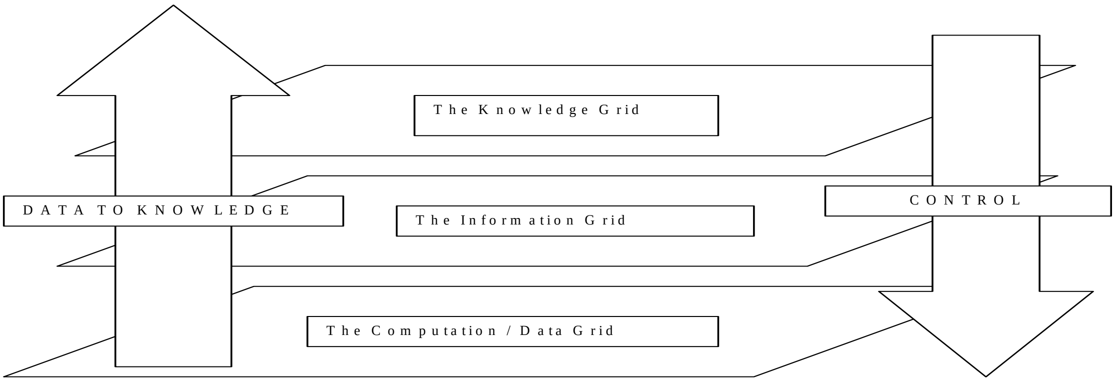

Figure 1: The 3-Layer GRIDs Architecture

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, 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.

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 (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.

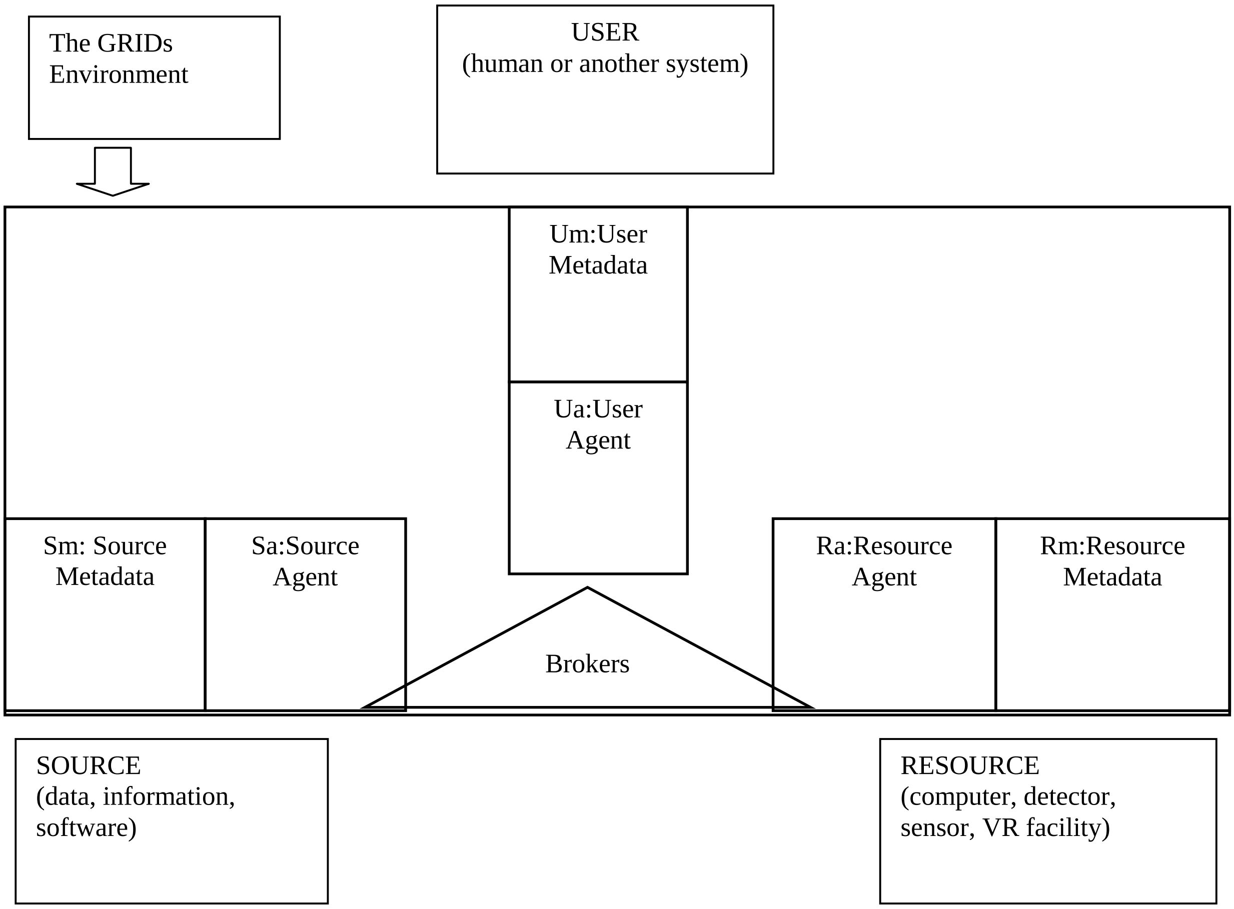

Figure 2: The GRIDs Components

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

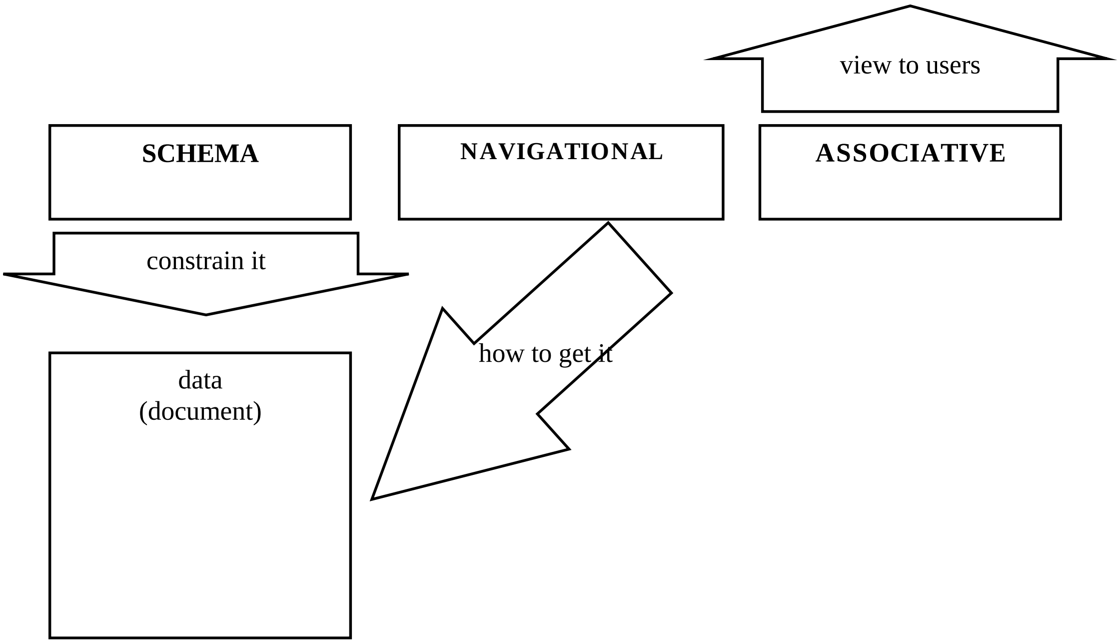

Figure 3: Metadata Classification

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.

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.

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.

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.

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.

Research Issues

The major issues are:

Metadata: Since metadata is critically important for interoperation and semantic understanding, there is a requirement for precise and formal representation of metadata to allow automated processing. Research is required into the metadata representation language expressivity in order to represent the entities user, source, resource. For example, the existing Dublin Core Metadata standard is machine-readable but not machine-understandable, and furthermore mixes navigational, associative descriptive and associative restrictive metadata. A formal version has been proposed (Jeffery, 1999).

Agents: Ths issue is generality or specificity of agents. Agents could be specialised for a particular task or generalised and configured dynamically for the task by metadata.

Brokers: Similarly, are they generalised and dynamic or specific? The degree of negotiation autonomy becomes the key research issue: can the broker decide by itself or does it solicit input from the external entity (user, source, resource) via its agent and metadata?

Security: Security is an issue in a common marketplace with great heterogeneity of purpose and intent. The security takes the forms: a) prevention of unauthorised access: this requires authentication of the user, authorisation of the user to access or use a source or resource and provision or denial of that access. The current heterogeneity of authentication and authorisation mechanisms provides many opportunities for deliberate or unwitting security exposure; b) ensuring availability of the source or resource: this requires techniques such as replication, mirroring and hot or warm failover. There are deep research issues in transactions and rollback/recovery and optimisation; c) ensuring continuity of service: this relates to (b) but includes additional fallback procedures and facilities and there are research issues concerning the optimal (cost-effective) assurance of continuity. In the case of interrupted communication there is a requirement for synchronisation of the end-user's view of the system between that which is required on the PDA and / or laptop and the servers. There are particular problems with wireless communications because of interception. Encryption of sensitive transmissions is available but there remain research issues concerning security assurance.

Privacy: the tradeoff of personal information provision for intelligent system reaction is the issue.

Trust: This concept requires much extension from current B2C e-Business to trust associated with e-tendering, e-contracts, e-payments, e-guarantees.

Conclusion

The GRIDs architecture will provide an IT infrastructure to revolutionise and expedite the way in which we do business and achieve leisure. The Ambient Computing architecture will revolutionise the way in which the IT infrastructure intersects with our lives, both professional and social. The two architectures in combination will provide the springboard for the greatest advances yet in Information Technology. The architecture is particularly relevant to field and laboratory geologists, to exploration geologists and geophysicists, to mining engineers and to academic geologists.

References

Jeffery, K.G., Gill, E.M., 1976 The Design Philosophy of the G-EXEC System. Computers and Geosciences 2(1976) 345-346.

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

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., Hutchinson, E. K., Kalmus, J. R., Wilson, M. D., Behrendt, W., Macnee, C. A., 1994, A Model for Heterogeneous Distributed Databases' Proceedings BNCOD12 July 1994; LNCS 826 pp 221-234 Springer-Verlag

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