Data and Metadata on the Semantic Grid

Geoffrey Fox
Community Grids Laboratory
Indiana University
gcf@indiana.edu

Introduction

Last issue, we covered the important role that the Grid is expected to have in information systems; roughly the Grid of Sensors and databases. We contrasted this with the major original Grid focus of metacomputing or the Grid linking distributed computers together. Whatever type of Grid one has, metadata or data about data will be important. The term is as always loosely defined and is often equivalent to information presented in "small high-value records". Examples of generally important Grid metadata include user information (name, address and the many profiles and preferences one accumulates) and specifications of the resources on the Grid. These could include for each computer the CPU, memory, and number of nodes (if parallel) while for software, there would location, compiler options and perhaps specification of needed input data. We have explained in previous articles, the underlying service model developed for both the Web and the Grid. This allows us to identify "service metadata" as a critical feature of the evolving architecture. Computers and software are both services in the Grid model but so also are databases, sensors, network and user information systems. In this article, we try to describe some of the many different sources and approaches to metadata.

Semantic Grid

Above we identified service metadata as essential to a Grid. In the current web service-based approach, each service exhibits input and output channels or ports which are each eventually specified in practice by a URL such as http://gridserviceNNN.niftywebsite.org:80XY. This emphasizes that web services are "just" web "pages" or better that accessing a web page is the most important example of a web service data. Thus studying metadata about web pages will be very helpful in understanding Grid metadata. This observation is at the heart of the Semantic Grid [1, 2] which builds on the important W3C Semantic Web [3, 4] initiative. The latter aims at more intelligent web searching and linkage based on attaching rich metadata to web pages. The well known article [5] by Berners-Lee and co-authors gives a nice health care example and suggests how metadata enriched web pages could allow one to identify an illness, a preferred treatment, optimal health care specialists and even the needed ambulance service. A more technical review [4] describes the Semantic Web as an approach to distributed artificial intelligence with tools to both generate the metadata attached to "web pages" (resources) and to reason about their significance especially when different resources are combined and the individual metadata can guide both what to combine and the significance of their combination. Combining health care web pages to generate an optimal diagnosis, doctor and ambulance services could be a rather loose coupling of text pages matching illness descriptions between doctor, diagnosis and patient as well as address information between patient and doctor. Combining services is more complex as one is matching rather precise input and output service ports corresponding to a message based implementation of remote procedure calls in the web or Grid service model. A web service specifies in its WSDL (Web Service Definition Language) instance the syntax of the procedure. The metadata adds to this the semantics or meaning of the methods in each service. Such enriched function (Grid service ports) calls is the characteristic of the Semantic Grid concept. Now we have services (programmed perhaps in C++ or Java) with an overlay of metadata arranged so that one can use it to deduce important consequences of linking services together. This combination of "distributed artificial intelligence" and conventional programming represents a new programming paradigm which could be very significant. The semantic or metadata information can be thought of as a richer form of the documentation traditionally attached to programs. This semantic information is built in terms of keywords (for example molecule, atom and binding energy in chemistry) and relationships (e.g. molecules are composed of atoms) which need to be developed by experts in each and every field. These domain specific semantic frameworks are termed ontologies.

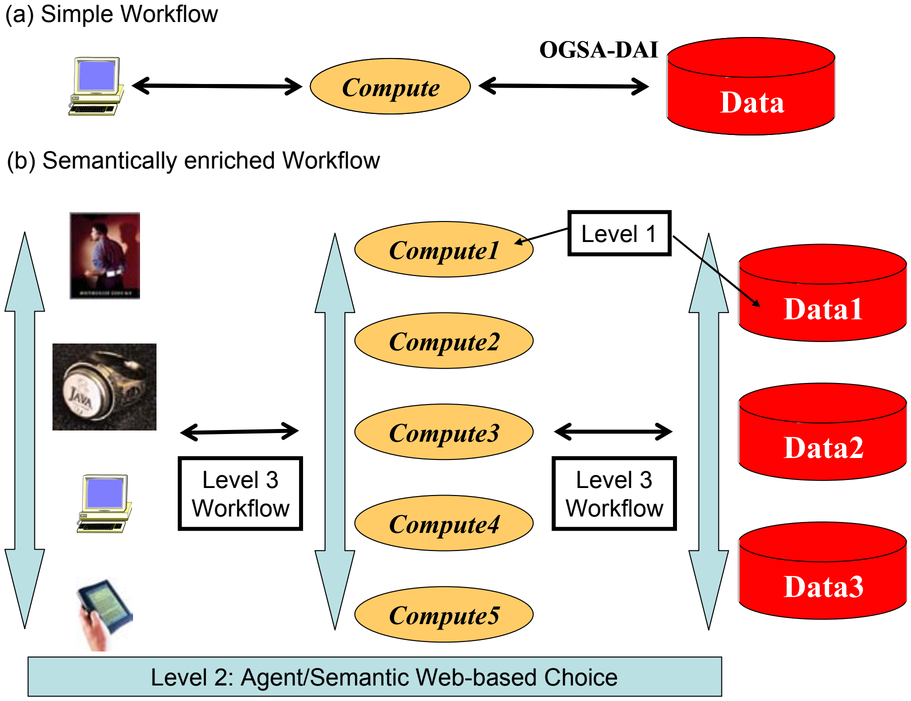

Fig. 1(a) A simple two-level workflow described in the last issue's column showing a pipeline linking database, compute and visualization resource. (b) The semantically rich three-level workflow adding agent (broker) based choice of resources at each stage of workflow.

Workflow and a three-level Programming Model

As briefly discussed in earlier articles, linking of services is usually called *workflow* in the Grid community and rich service metadata is usually considered essential for any dynamic workflow where real-time decisions are being made on which services to tie together to solve a particular problem. Note one often imagines there will be many contenders to provide a given service. As an example, consider a bioinformatician who needs to extract gene data from one or more databases and send it to a computer to be matched together. One might use metadata to choose the database source (the metadata could describe your view of the quality of the database), the computer resource (whose dynamic metadata includes its cost and availability) and visualization engine (whose metadata could include functionality and suitability for client display). One might use "agents" to match metadata between resources and this leads to a three level programming model illustrated in fig. 1.The lowest level is the traditional code (SQL, Fortran, C++ or Java) implementing a service; the next level is agent technology using metadata to choose which services to use; the third workflow level links the chosen services to solve the distributed bioinformatics problem. Each level has distinct "programming models" with the metadata of relevance here needing tools to add (annotate) the metadata and some logic (artificial intelligence) to reason about them. These tools have initially been developed by the Semantic web community and are now being applied to the Grid. MyGrid and DiscoveryNet are two UK e-Science projects exploring these ideas for Bioinformatics. Note this discussion has shown how the Grid requires and can use both distributed artificial intelligence (AI) and agent technologies and the three level programming model integrates "conventional programming", AI, agents and coarse grain application integration or workflow; four often disparate and rival threads of computer science research.

Usually one considers Information or Data Grids as providing a pathway driving data to information and thence to knowledge; the process of fig. 1 shows the paradigm by which we do this. In this regard, metadata is rather loosely defined as it includes metadata, meta-information and meta-knowledge. In the following sections we elaborate on the many types of metadata, their search and retrieval and how they can be gathered. These are but the briefest of discussions which is expended with many references in our review [6].

Types of Metadata

Above we have discussed a few forms of metadata including:

a) Registry giving URL's (locations) in terms of URI's (unique identifiers) for each Grid or web resource.
b) From the resource view, basic metadata specifying computers databases and users
c) From the Web Services view, basic listings and simple look-up of available services relating function and location without significant rich context.
d) From the Semantic Grid view, sophisticated ontologies and metadata to provide intelligent lookup and matching of services

However there are many other related types of "small data" which share either support technologies or use with service and resource metadata given above. These include [6]:

e) From the monitoring view, streams of information recording the operation of Grid Resources. This would include both network performance and job status data.
f) From the collaboration and caching point of view, state changes in services to support coherence between different copies and versions of the same basic resource.

g) From each application field, specialized metadata to manage the data and information in each field. These would be typically be stored in metadata catalogs maintained in a manner and format idiosyncratic to each application.

h) From the provenance and data curation view, metadata describing life-cycle and ownership of data. This area linking Grid and digital library communities has been highlighted as an topic needing greater attention.

Gathering of Metadata

One of the major issues is how to generate metadata; using an analogy above, we know how hard it is to document software and processes and so we can expect metadata generation to be challenging in general. The Semantic web project has generated several annotation tools allowing convenient interfaces for users to add metadata describing resources. Curation as is common in Bioinformatics where one sets up teams to manage databases and both to insure the quality of information added and to generate some of the needed metadata; many scientific fields have set up such processes with specific organizations responsible among other things for metadata generation.

Much metadata will be generated "automatically" from data or other services connected to a particular resource. Some computer will use existing metadata and data to produce new important metadata. One simple example is provenance information defining the processing and ownership of data which corresponds to some extent to preservation of computer generated auditing metadata. More controversially, we can use search engines like Google to "automatically" find metadata by comprehensive sieving through the complete datasets.

Storage Search and Access to Metadata

Treatment of storage, search and retrieval of metadata in the Grid could be dismissed by noting that it is "just a database problem" and refer to last issue's article with the discussion of OGSA-DAI to integrate databases with the Grid. This is basically the required approach as long as we take a liberal view of a database as stretching from a flat file to the latest release of Oracle. However fig. 2 emphasizes a key difficulty as metadata is intrinsically distributed with some (the key service registry data) being held in a system wide database; this could in fact be formed by the federation of several different individual repositories. On the other hand, some metadata will also be stored in the service itself; this is exemplified by Service Data elements (SDE) in the new OGSI Grid service model [7] or by use of embedded metadata tags in an HTML web page. In OGSI, information ports allow one to query and retrieve such embedded metadata. The middle layer of fig. 2 shows an intermediate situation where some metadata is stored in Grid or domain specific databases with a limited scope nd often using very different technologies and policies. For any given service, one expects some metadata to be stored internally but other only in external repositories; for example as one adds metadata to the world's web pages, it is usually not practical to add this to the page itself – it must be placed external to the resource it describes. Thus managing metadata requires attention to distributed repositories with very variable scope and completeness. Further a repository at one level may be (partially) generated by the injection of more distributed metadata at a lower layer in fig. 2. This complex situation has no complete solution at present and we can expect substantial progress in the near future as more sophisticated database technology is deployed.

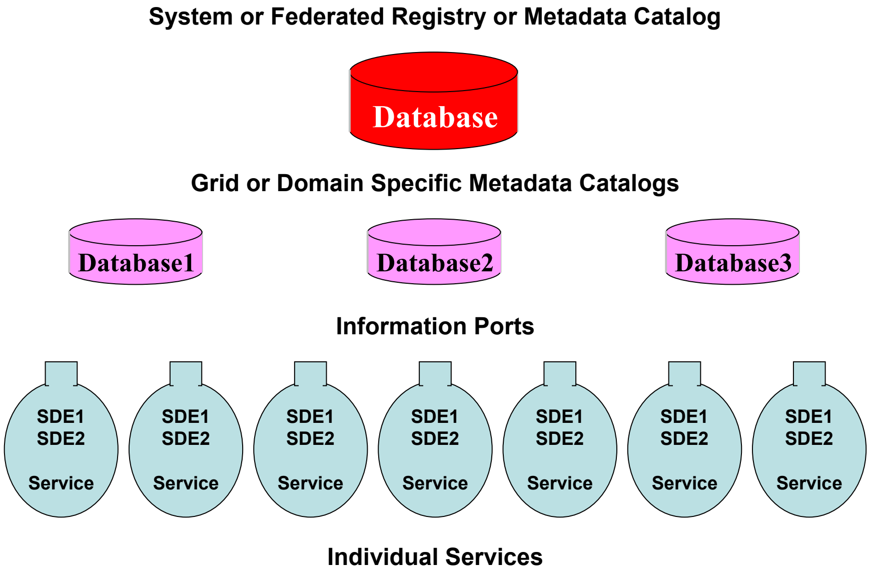
System or Federated Registry or Metadata Catalog
Database
Grid or Domain Specific Metadata Catalogs
Database1
Database2
Database3
Information Ports
SDE1
SDE2
Service
Individual Services

Fig. 2: Hierarchy of distributed Metadata repositories

Conclusion

Metadata is essential for the Grid and suggests new programming models combining workflow, conventional languages and distributed artificial intelligence; this is the Semantic Grid vision. The core database technology to generate, store, search and access metadata must address difficult federation and distributed system issues. We are still far from agreed mature solutions to these issues. We gave a brief review of these latter issues which should be followed up by the reader interested in more detail.

Further Reading

1) http://www.sematicgrid.org

2) *The Semantic Grid: A Future e-Science Infrastructure* :David De Roure, Nicholas Jennings and Nigel Shadbolt, Chapter 17 of *Grid Computing: Making the Global Infrastructure a Reality* edited by Fran Berman, Geoffrey Fox and Tony Hey, John Wiley & Sons, Chichester, England, ISBN 0-470-85319-0, March 2003. http://www.grid2002.org

3) http://www.w3.org/2001/sw/

4) IEEE Intelligent Systems March/April 2001 pages 24-79, *Semantic Web Issue* http://www.computer.org/intelligent/ex2001/x2toc.htm

5) Berners-Lee, T., Hendler, J., and Lassila, O., *The Semantic Web*, Scientific American, May2001.

6) Section 7.6 of *e-Science Gap Analysis*, by Geoffrey Fox and David Walker, report UKeS-2003-0, http://www.nesc.ac.uk/technical_papers/UKeS-2003-01/index.html

7) OGSI Open Grid Service Infrastructure Working Group of Global Grid Forum http://www.gridforum.org/ogsi-wg/