# Report on the Workshop on Metadata Management in Grid and Peer-to-Peer Systems, London, December 16 2003

Kevin Keenoy[1], Alexandra Poulovassilis[1], Vassilis Christophides[2], George Loizou[1]
Giorgos Kokkinidis[2], George Samaras[3], Nicolas Spyratos[4],

[1] School of Computer Science and Information Systems, Birkbeck, University of London

[2] ICS-FORTH, Heraklion, Crete

[3] Department of Computer Science, University of Cyprus

[4] Laboratoire de Recherche en Informatique, Université Paris-Sud

## 1 Introduction

A workshop on *Metadata Management in Grid and Peer-to-Peer Systems* was held in the Senate House of the University of London on December 16 2003.[1] The workshop was organised by the *SeLeNe* (Self e-Learning Networks) IST project as part of its dissemination activities.[2] The goal of the workshop was to identify recent technological achievements and open challenges regarding metadata management in novel applications requiring peer-to-peer information management in a distributed or Grid setting. The target audience for this event were researchers from the Grid, peer-to-peer and e-learning communities, as well as other application areas requiring Grid and/or peer-to-peer support. The event attracted 43 participants from 8 different European countries, and we believe that it was an important step in coordinating research activities in these inter-related areas. The presentations at the workshop fell into one of four sessions, each of which we report on below.

[1] All presentations from the workshop are available from the workshop website at `www.ics.forth.gr/isl/ist_workshop/` and are also published on-line in a repository for workshop proceedings at `http:CEUR-WS.org`.

[2] See `www.dcs.bbk.ac.uk/selene/` for a description of the SeLeNe project and its key transferable outcomes, publications, and other related resources.

## Session 1: Metadata Management in Grid systems

This session contained three presentations that discussed the metadata present in Grid systems, how Grid services can be provided using existing technology, and how metadata can be used in new Grid domains. In all three presentations the need for metadata management middleware was evident.

The first presentation, by Gavin McCance, presented how DataGrid, the European flagship Grid project, addresses issues of metadata management. DataGrid is a file-based system manipulating quantum-physical, Earth-observation and biomedical data. The DataGrid project utilises two types of metadata: Grid-internal metadata and application-specific metadata. The system supports a Replica Metadata Catalog and a Replica Location Service. It uses bloom filters and supports a flexible search facility. There is as yet no generic application metadata management middleware and it is up to the end-user to provide the specific application metadata and to request the relevant files to manipulate.

The second presentation by Savas Parastatidis focussed on how Grid services can access an organisation's resources. He identified the need for resource metadata that can be published outside the organisation's boundaries and argued that this can be achieved via existing specifications and tools

without the need to change the existing infrastructure. He presented two possible solutions, one based on Service Data Elements (SDEs) and one based on the Grid Resource Metadata document (GRM). The advantage of the latter is that the GRM is based on XML Schema and therefore no additions to WSDL or any other specifications are needed, existing tools work, it does not add semantics to a Web service (as it is just a document) and it could be published in a registry. The GRM document provides a functional equivalent to SDEs that allows the re-use of existing technologies and work, which is not the case with SDEs.

In the third presentation, by Mario Cannataro, the 'Knowledge Grid' was presented as an environment that integrates data-mining techniques and Grid resources to build Grid-aware data-mining applications. The claim was that metadata describing data mining tools, data sources, Grid resources and ontologies for semantic modelling of the application domain is a must if the Grid is to be utilised in knowledge management and data mining. The need for two types of metadata was identified, namely for data mining tools and for classical data source metadata. The presentation showed how metadata and ontologies can be used to build and execute distributed data mining applications on the Knowledge Grid.

## Session 2: Metadata Management in P2P systems

The first presentation in this session was by Henrik Nottelmann, presenting work carried out jointly with Norbert Fuhr. The talk focussed on the use of probabilistic logics for defining and using service descriptions in a peer-to-peer network with a large number of web services. The goal of this work is to dynamically compute execution plans for services required to implement a given task. In this respect, the DAML-S upper ontology is used for describing services, while probabilistic Datalog is used for match-making. A service is described in DAML+OIL by its profile, which is used for match-making, by its model, which describes how the service works, and by its grounding, which describes how to access the given service. An ontology for library services (e.g. search, schema mapping, or result modification services) is also introduced, containing definitions for generic search services and other query- and result-transformation services. Match-making rules use facts derived from DAML-S to provide execution plans of services. Probabilities are used as a simple kind of cost estimation for plans.

The next talk by Philippe Cudré-Mauroux, on work carried out with Karl Aberer, presented the Chatty Web approach for global semantic agreements. The posed problem is that of achieving semantic interoperability among heterogeneous data sources in a peer-to-peer data management system, without relying on pre-existing global semantic models. The solution involves the use of local translations that enable global agreements. Semantic 'gossiping' is used for query forwarding and distribution through the system. An analysis is done between original and transformed queries based on syntactic and semantic distance measures. For the semantic similarity, query cycles can be detected and analysis of the results with content-retrieval techniques can be used. Consequently, a self-repairing semantic network can be organised with the use of evaluations based on Chatty Web simulations and the automatic correction of erroneous mappings based on gathered evidence.

The third presentation, by Wolfgang Nejdl, addressed issues related to data-centric networks and peer-to-peer data management. The speaker presented the main results and challenges of several ongoing and past projects. REWERSE is an EU project addressing retrieval, protection and integration of data. EDUTELLA has specified and implemented an RDF-based metadata infrastructure for peer-to-peer data networks. PROLEARN is working towards innovative and interoperable e-learning resources and sustainable e-learning infrastructures and processes. The notion of schema and the use of RDF/S for describing distributed resources were identified in these projects as useful parts of a peer-to-peer data management system. RDF-QEL is a

Datalog-based Query Exchange Language that can be used to wrap other RDF and XML query languages. The HyperCup peer-to-peer topology and its broadcast algorithm were discussed as a solution to the routing problem in a peer-to-peer system. Access control and automated trust negotiation were also identified as an important problem. All the above issues are essential building blocks for forthcoming schema-based peer-to-peer networks and peer-to-peer-based data management infrastructures.

## Session 3: Applications I

The first presentation in this session, from Tamás Hauer, concerned the role of metadata in querying Grid-resident medical images in the MammoGrid project, an EU-funded project aiming at a Grid solution for mammography and involving three hospitals among its partners. A federated system solution is proposed whereby, with the help of shared metadata, a clinician can address a query to the system, which is then translated to remote local sub-queries whose results are then returned to the clinician. The specific characteristics of the medical domain necessitate flexibility and extensibility as well as management of domain information such as annotations of images. These and other considerations have led to a service-oriented architecture using Grid-middleware. Each participating node is responsible for managing its own metadata and can change its service description on the fly, new sites can join seamlessly, and domain and service ontologies are defined independently.

The second presentation, by Bob Bentley, concerned metadata management in virtual solar observatories and reported experiences from the EGSO project, a Grid test-bed designed to improve access to solar data for the solar physics and other communities. The EGSO project addresses the generic problem of a distributed heterogeneous data set and a scattered user community. In such a setting, the availability of good quality metadata is important for searching. Resources are described by entries in a resource registry and managed by a broker. Brokers and registries are replicated to provide system resilience and to enable load sharing.

The third presentation, by Theo Dimitrakos, concerned metadata management issues underpinning emerging solutions for distributed trust and contract management and enforcement in enterprise Grid and peer-to-peer systems. He presented in some detail four projects where metadata play an important role: GRASP, Grid-based application service provision; CORAS, a CASE tool and method support for security-risk analysis; SWAD-Europe, Semantic Web technology development; and PEL-LUCID, an agent-based platform supporting organisational mobility.

The final presentation, by Stavros Christodoulakis, discussed metadata management for audiovisual content to support intelligent video-content retrieval and e-learning services in digital TV (t-learning). These objectives are pursued in the context of the TV-Anytime framework, MPEG-7 and SCORM. An ontology-driven framework for semantic metadata management was presented. In this framework, TV-Anytime keywords are used to describe program segments. Domain specific ontoloties are built using the MPEG-7 semantic model. Ontologies are used for filtering and retrieval of MPEG-7 multimedia content. Semantic annotations are transformed into TV-Anytime segment keywords. With respect to t-learning, two objectives were identified: providing interoperability for educational applications in different e-learning and digital TV environments; and creation of metadata for digital TV for educational purposes in order to offer educational experiences exploiting TV programs.

## Session 4: Applications II

The fourth session of the Workshop consisted of four presentations on e-learning applications that rely on some form of schema-based repository of Learning Object (LO) metadata.

The first presentation, by Klaus Jantke, gave an outline of the DaMiT e-learning system developed to provide a tutorial on data mining. The system allows users to specify several of their pref-

erences as a 'profile', including the type of material preferred (example-oriented, theory-oriented, etc.), style (formal, informal) and language (although the system is mostly German-only at present). The e-learning content delivered by the system is stored as a collection of fragments of learning material, each described by associated metadata. DaMiT generates complete pages of learning material from these fragments on-the-fly, adapting the page to the user's needs based on their profile. DaMiT also includes an e-payment system that uses metadata about the user to keep track of their access authorizations to the learning material.

The second presentation was by Kevin Keenoy on the SeLeNe project. SeLeNe has developed techniques enabling the discovery, sharing and collaborative creation of LOs. LOs to be shared within a learning community are described using an extended form of the IEEE-LOM metadata schema, encoded in RDF, and these descriptions form a distributed repository of metadata that can be queried by users searching for learning material. The metadata entries make use of taxonomies of subject and topic domains, learning objectives and learning styles. User profiles are also encoded using RDF, and make use of the same taxonomies as the LO descriptions. The novel features of SeLeNe include: registration of composite LOs, dependent on other atomic or composite LOs; automatic generation of taxonomical information for composite LOs; definition of personalised views over LO descriptions and schemas using the RVL language; generation of personalised query results based on the user profile; and personalised event and change notification services by means of event-condition-action rules on RDF.

The third presentation, by David Massart on work carried out jointly with Frans Van Assche, described research in the Celebrate project on developing a system for metadata search and exchange of LOs, creating a European Learning Network (ELN) that is used by 500 schools across Europe. The ELN is built around a brokerage system that manages exchanges between its members, enabling interoperability between e-learning systems by searching and exchanging LOs in their repositories. The IEEE-LOM schema is used for LO descriptions. Various communication protocols may be used between different ELN members, but the individual systems are shielded from this complexity by the 'ELN client' that provides a simple communication API. A brokerage system hosts a central metadata repository, and each ELN member is also authorised to manage its own local metadata repository. Search requests are both handled centrally and propagated to local repositories.

The final presentation, by Zoltán Miklós on work carried out jointly with Bernd Simon, described research in the ELENA project on "creating a smart space for learning". ELENA, like SeLeNe and Celebrate, is concerned with the sharing of learning resources that are contained in heterogeneous, distributed repositories. Metadata repositories are connected in a peer-to-peer fashion, and Semantic Web techniques are used to achieve interoperability. Personal Learning Assistants support learners in selecting appropriate learning services from the available sources. The metadata for learning services is based on a common formal ontology. Learners can define views of the ontology using the TRIPLE language. A Simple Query Interface translates queries into an appropriate form, establishing interoperability among heterogeneous learning repositories.

The presentations in this session identified several common issues arising in systems providing personalised e-learning. In systems that employ a user profile two main issues are content and representation — what information is needed about the user, and in what format should it be stored? There is also the issue of where from the information for the profile comes — is it user-supplied, teacher-supplied, automatically derived, or some combination of these three? A further issue is matching the profile with LO metadata — how can the profile be used to provide the most useful personalised LOs or personalised ranking? Future work can empirically test different matching algorithms in different learning situations since present methods generally seem to be sensible ad-hoc choices rather than empirically validated solutions. Another area requiring future investigation is that of learning style taxonomies — which ones are most useful in the con-

text of e-learning, and are what the relationships among different learning style taxonomies.

## Conclusions

During the discussion session at the end of the workshop it was generally agreed that this workshop had been very beneficial and timely in bringing together the common strands of research from the Grid, P2P and applications communities. Three sets of research challenges emerged in the area of metadata management for Grid and peer-to-peer systems from the presentations and discussions at the workshop:

1. Challenges stemming from the *distribution*, *autonomy* and *heterogeneity* of information and services, including the need for: metadata describing information and services available at the nodes of a Grid/P2P system; searching and matching techniques utilising this metadata for discovery of information and services; metadata and techniques for controlling access to and privacy of information and services; techniques for automatic service composition and orchestration; metadata and techniques for translation of queries and query results in the absence of a controlled global schema; metadata supporting replication and consistency of information; automatic/semi-automatic extraction of metadata from large of heterogeneous data.

2. Challenges arising from the *dynamicity* of Grid/P2P environments, e.g. changes in the network topology, information content at nodes, and service availability at nodes. This leads to the need for the techniques being developed under area 1 above to be scalable, adaptive, extensible and fault-tolerant.

3. Challenges arising from the *heterogeneity of users* accessing Grid/P2P systems. This leads to the need for techniques that personalise content and presentation to different users' needs and preferences.

During the discussion at the end of the workshop it was generally agreed a similar event in a year's time should be aimed for. We are therefore pleased to announce the *2nd Workshop on Metadata Management in Grid and P2P Systems* to be held on 17th December 2004, Senate House, University of London.