INFORMATION QUALITY EVALUATION FOR GRID INFORMATION SERVICES

Wei Xing, Oscar Corcho, Carole Goble
School of Computer Science
University of Manchester
United Kingdom
wxing@cs.man.ac.uk
ocorcho@cs.man.ac.uk
carole@cs.man.ac.uk

Marios Dikaiakos
Department of Computer Science
University of Cyprus, Cyprus
mdd@cs.ucy.ac.cy

Abstract The quality of the information provided by information services deployed in the EGEE production testbed differs from one system to another. Under the same conditions, the answers provided for the same query by different information services can be different. Developers of these services and of other services that are based on them must be aware of this fact and understand the capabilities and limitations of each information service in order to make appropriate decisions about which and how to use a specific information service. This paper proposes an evaluation framework for these information services and uses it to evaluate two deployed information services (BDII and RGMA) and one prototype that is under development (ActOn). We think that these experiments and their results can be helpful for information service developers, who can use them as a benchmark suite, and for developers of information-intensive applications that make use of these services.

Keywords: Grid, Grid Information Service, Information Quality, Evaluation

1. Introduction and Motivation

Information Services are regarded as a vital component of Grid infrastructure. They address the challenging problem of discovery and monitoring of a variety of Grid resources, including services, hardware, software, etc. The quality of information provided by information systems affects the performance and the behaviour of other dependent Grid services. For instance, a Grid meta-scheduling service will not work optimally if the quality of the information used for decision making is poor; a Grid Resource Broker depends on the Grid resource information provided by the information services that it uses; etc.

There is little work on the evaluation of information quality of Grid information services. Most evaluation studies focus on performance measurement [1], evaluating scalability, overload, query response time, etc. Such measurements are based on the assumption that information quality is equal for different information services. However, this assumption does not hold in reality, since each information system has different mechanisms for collecting and processing information, and adopts different information models for storage and querying. We cover this in our experiments, which show that even for a simple query each system provides different results. For example, for the query "*find me Computing Elements which support the Biomed Virtual Organisation*" the two EGEE default information services, BDII and RGMA, gave 151 and 30 results respectively. Independently of the reasons for such differences, the main outcome from this simple test is that information quality of currently-deployed Grid information services has to be considered carefully.

The work described in this paper has several objectives. First, we want to obtain a **fair systematic approach to measure information quality of different Grid information services**, so that we can compare them and provide guidelines related to when each of them can be used. One challenge is related to the fact that different Grid information services have different information models to represent the same type of Grid resources: some of them use LDAP to represent that information and others use relational models, and the information that they store about each resource may also differ. Unlike information quality evaluation in other domains (such as Web search, where precision and recall measurements can be obtained by counting numbers of documents), the information objects in our evaluation are heterogeneous, both in the information model used and in its access API, what makes it hard to compare the outputs. We have proposed the use of a common information model to allow comparisons between these outputs, as explained in Section 3.2. Another challenge is related to the differences in the querying capabilities and expressiveness supported by each service, what makes it difficult to design a good set of relevant experiments for the evaluation. Some services allow making complex queries that relate information from different domains (*computing elements that sup-*

port a specific virtual organisation and a specific software environment) and others just provide simple querying functionalities. In our approach we have proposed a set of representative queries that may be issued by other middleware services or applications, with increasing levels of complexity.

Our second objective is to use our approach to **evaluate information quality of two EGEE information services (BDII and RGMA) and one prototype that is under development (ActOn)**. We will analyse the results from this evaluation and identify the reasons for obtaining them. These results can be used by developers working on these Grid information services, in order to improve them, and by developer of systems that are based on them.

The remaining of this paper is organised as follows. Section 2 describes the information systems to be evaluated. Section 3 introduces our evaluation framework, including the design rationale, the experiments, and the metrics to use for evaluation, together with details about how they are measured for each system. Section 4 describes the results of the experiments carried out, and provides some conclusions related to these results. Finally, Section 5 reflects about the lessons learnt in the design of this evaluation framework and gives references to additional performance tests that we have carried out.

2. Grid Information Services

Currently, there are several well-known and widely-used Grid information services: Monitoring and Discovery System (MDS), Berkeley DB Information Index (BDII), and RGMA [2–3]. These services are deployed in most Grid systems, such as Europe Data Grid, Crossgrid, and Open Science Grid, and widely used by Grid middleware and applications running on them. From these three services, we will select BDII and RGMA for our evaluation, since they are the default information services for the EGEE Grid. We do not include MDS because it is not used for Computing Elements (CEs) and Sites in EGEE and would make difficult to perform the comparison. Besides, BDII is based on MDS, with the same information model (information representation and access), hence the general results regarding information quality and recommendations obtained for BDII could be easily extrapolated to MDS. Besides these two services, we will evaluate our ontology-based information service (based on the ActOn [4] ontology-based integration architecture).

Berkeley DB Information Index (BDII) [2] is an improvement of MDS, the information service component of the Globus platform. It uses the MDS information model and access API and caches information with the Berkeley DB. Information about Grid resources is extracted by "information providers", software programs that collect and organise information from individual Grid entities, either by executing local operations or by contacting third-party information sources.

Relational Grid Monitoring Architecture (RGMA) [3] combines monitoring and information services based on a relational model, implemented with XML. It has been built in the context of the EU DataGrid project and implements the Grid Monitoring Architecture (GMA) proposed by the Open Grid Forum. GMA models the Grid information infrastructure with three types of components: information producers, information consumers, and a registry, which mediates the communication between them.

Active Ontology (ActOn)-based information service ActOn [4] is an ontology-based information integration system, developed by us, which can be used to maintain up-to-date information for dynamic, large-scale distributed systems. The ActOn architecture is comprised of a set of knowledge components, which represent knowledge from the application domain (e.g., the EGEE Grid) and from the information sources (e.g., RGMA and BDII servers); and software components, such as a metadata scheduler (MSch), an information source selector (ISS), a metadata cache (MC), and a set of information wrappers.

We will evaluate a deployment of the ActOn system that uses BDII and RGMA as information sources, and a Grid Ontology [5–6] as its information model, and has been deployed in the EGEE certificate and production testbeds.

3. An evaluation framework for information quality in Grid information services

Information quality (IQ) can be defined as a measure of the value of the information provided by an information system to its users [7]. Quality is normally subjective and depends on the intended use of information. The authors in [7] distinguish a set of quality features (intrinsic, contextual, representational and accessibility IQ) and define different factors to be considered for each of them (accuracy, objectivity, reputation, relevancy, etc.).

The authors in [8] propose to focus on seven of these characteristics: completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility. We have selected three of them, namely *completeness*, *accuracy* and *conformance to expectations*.

We are not worried about the *provenance* of information, since we know clearly which are the information sources that we use in each moment and which are the information providers responsible for that information. We are not worried either about *accessibility*, since we assume that the systems work within a Grid security infrastructure (e.g., GSI), so that the information is accessible as long as the client has the rights to access it and knows the information model and API used by the corresponding information service.

With respect to the *logical consistency and coherence* and the *timeliness* of the information retrieved and aggregated from the information sources, these are features that will form part of our future evaluation work, and will be also considered in further developments of the ActOn-based information service. An example of why the first feature is important is the following: there are many cases where a computing element specifies that it gives support to MPI but does not comply with the requirements for running an MPI job, which are that it must be a CE server, must have an `sshd` service running on it, must have the libraries `mpirun` and `libmpi.so` in its file system, and must have at least two worker nodes. Information services like BDII or RGMA only store and provide the information that their information producers give them, without checking their consistency, hence they provide incorrect information due to this fact. As an example of the second feature, BDII normally updates the information that has been provided by its information sources every five or six minutes, what means that this information may be already inaccurate when a client requests it. Hence, having metadata about the lifetime and freshness of information in the information service is important.

Now we describe our information quality evaluation framework, including metrics to be used, the design rationale, and the experiments, together with details about how the metrics are obtained for each system.

3.1 Evaluation metrics

To check our three criteria, we want to know whether all information services obtain the same results when answering the same query, given the same conditions in the EGEE testbed. We also want to check how many of those answers are correct and how many of the existing answers are actually retrieved. This allows us to know whether the results provided by the services conform to the user expectations. To check this, we have selected two metrics commonly used in information retrieval: precision (The proportion of relevant information retrieved, out of all the information retrieved) and recall (the proportion of relevant information that is retrieved, out of all the relevant information available).

3.2 Experiment setup and design

Measurements are taken on the EGEE production testbed, which are accessed through the UI machines at the University of Manchester[1] and at the Institute of Physics of Belgrade[2]. A set of Java-based client software and Unix shell scripts have been developed to carry out the experiments and record their results. They are available at [6].

The key aspects upon which we compare different information services are their information model and the expressiveness of their query language. To

[1] ui.tier2.hep.manchester.ac.uk

[2] ce.phy.bg.ac.yu

evaluate these two features, we have proposed six representative queries that cover a wide range of Grid systems (hardware and software resources, middleware environment, services, applications, etc.) with increasing complexity:

- Query 1: Find all the Computing Elements (CEs) that support the BIOMED Virtual Organisation (VO).
- Query 2: Find all the CEs that support the BIOMED VO and have more than 100 CPUs available.
- Query 3: Find all the CEs that support the MPI running environment.
- Query 4: Find all the CEs that support the BIOMED VO, have more than 100 CPUs available, and support the MPI running environment.
- Query 5: Find all the CEs where GATE (Geant4 Application for Tomographic Emission) can be run.
- Query 6: Find all the CEs that support the BIOMED VO, have more than 100 CPUs available, and where GATE can be run.

Table 1. An Example of the Query 1 in BDII, RGMA, and ActOn

Information Service	Query 1

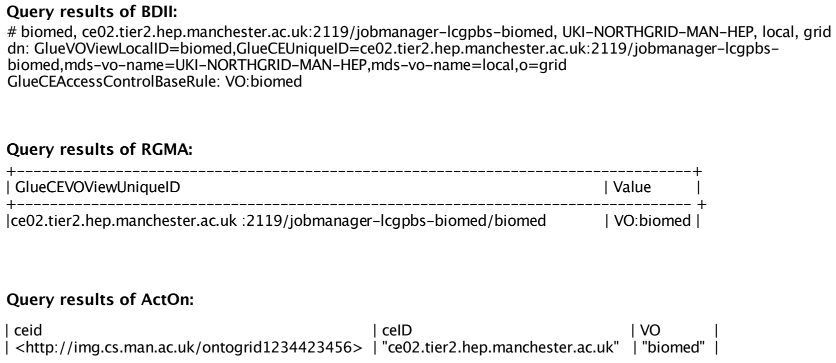
BDII (LDAP Search)	`ldapsearch -x -H ldap://lcg-bdii.cern.ch:2170`
	`-b mds-vo-name=local,o=grid ’(&(objectClass=GlueVOView)`
	`(GlueVOViewLocalID=biomed))’ GlueCEAccessControlBaseRule`

RGMA (SQL Query)	`select GlueCEVOViewUniqueID, Value from`
	`GlueCEVOViewAccessControlBaseRule WHERE Value=’VO:biomed’`

ActOn (SPARQL Query)	`PREFIX egeeOnto: <http://www.cs.man.ac.uk/img/ontogrid#>`
	`SELECT ?ceid ?ceID ?VO`
	`WHERE`
	`?ceid egeeOnto:CEUniqueID ?ceID .`
	`?ceid egeeOnto:hasVO ?VO .`
	`OPTIONAL { ?ceid egeeOnto:VO ?ceID .`
	`FILTER ( ?vo = ‘‘biomed’’)}`

Each query has been translated into the query languages of the three information services. Table 1 shows an example for Query1. We use different clients to execute them and extract the results (e.g., ldapsearch for BDII, the gLite RGMA client tools for RGMA and a Java-based ActOn client for the ActOn-based information service).

Results are obtained in different manners. The result of a BDII query is a set of LDAP entries, of an RGMA query a set of table rows, and of an ActOn-based query a set of RDF triples. Figure 1 shows three different ways to show the same Grid resource (ce02.tier2.hep.manchester.ac.uk, an EGEE Computing Element) in the three services evaluated. In our experiment we use each "Grid resource" obtained from a query as the basic unit for counting information, which will be used to calculate precision and recall.



```
Query results of BDII:
# biomed, ce02.tier2.hep.manchester.ac.uk:2119/jobmanager-lcgpbs-biomed, UKI-NORTHGRID-MAN-HEP, local, grid
dn: GlueVOViewLocalID=biomed,GlueCEUniqueID=ce02.tier2.hep.manchester.ac.uk:2119/jobmanager-lcgpbs-
biomed,mds-vo-name=UKI-NORTHGRID-MAN-HEP,mds-vo-name=local,o=grid
GlueCEAccessControlBaseRule: VO:biomed


Query results of RGMA:
+-------------------------------------------------------------------------------------+
| GlueCEVOViewUniqueID                                                  | Value      |
+------------------------------------------------------------------------------------- +
|ce02.tier2.hep.manchester.ac.uk :2119/jobmanager-lcgpbs-biomed/biomed   | VO:biomed |


Query results of ActOn:

| ceid                                         | ceID                                | VO        |
| <http://img.cs.man.ac.uk/ontogrid1234423456>  | "ce02.tier2.hep.manchester.ac.uk"   | "biomed"  |
```

Figure 1. Results of BDII, RGMA, and ActOn for the the same Grid resource Computing Element at University of Manchester (ce02.manchester.ac.uk)

3.3 Experimental Results Measurement

In the experiment we examine the information retrieved for each of the six queries, so as to get their corresponding precision and recall measures.

Precision is easy to determine, since it can be computed manually by looking at the results obtained from each query. In all cases, we assume binary relevancy of information, that is, each piece of information retrieved is either relevant or irrelevant for the issued query.

Recall is more difficult to determine, due to the fact that the amount of information available in the EGEE production testbed changes frequently in these systems and there is no way to get accurate information about the actual state of the Grid resources that are available without using the information services that we are evaluating. To get a good approximation that can be used for our purposes, we execute each query 100 times, with a 4-minute interval between executions, that is, we monitor the testbed during 400 minutes. Then we use the highest value obtained from this 100 executions as the total number of relevant information to be used to calculate recall.

4. Evaluation Results and Conclusions

Tables 2, 3 and 4 provide the precision and recall measurements obtained after the execution of the previous experiments for the three information services: BDII, RGMA and the ActOn-based information service. The values in the tables show the average of executing the queries 100 times.

Table 2. BDII Recall & Precision Measurement (100 times)

$QueryNo.$	$Retrieved\ Info.$	$Relevant\ Info.$	$Precision$	Recall
1	14,999	15,200	1	0.987
2	242,517	19,708	0.082	0.918
3	7174	7300	1	0.983
4	485034	4600	0.010	0.990
5	-	-	-	-
6	-	-	-	-

Table 3. RGMA Recall & Precision Measurement (100 times)

$QueryNo.$	$Retrieved\ Info.$	$Relevant\ Info.$	$Precision$	Recall
1	3417	15200	1	0.225
2	6321	6321	1	1
3	6568	7300	1	0.900
4	11245	4914	0.437	0.563
5	-	-	-	-
6	-	-	-	-

Table 4. ActOn Recall & Precision Measurement (100 times)

$QueryNo.$	$Retrieved\ Info.$	$Relevant\ Info.$	$Precision$	Recall
1	15200	15200	1	1
2	34100	34100	1	1
3	6568	7300	1	0.900
4	6568	7300	1	0.900
5	24	24	1	0.900
6	6	6	1	1

As a general comment, we can highlight the fact that BDII shows in general poor results with respect to recall and precision, while ActOn and RGMA present better results. This is mainly related to the repository that BDII uses (LDAP), which is too lightweight and hence provides weak information process and query capabilities; while RGMA's is based on relational databases and ActOn's is based on RDF, which both have better query capabilities.

Now we will analyse with more detail some of the system behaviours over specific queries, and derive more conclusions from these values:

BDII has weak query capabilities. Table 2 shows bad precision results for BDII in queries 2 and 4, while the results for queries 1 and 3 are excellent. This is related to its weak query ability. LDAP-based queries are string-based, and hence they cannot support queries over numerical values, such as "greater than or lower than". To improve this precision value, we need to fetch all information about CE CPUs as a string value first (as we have done to get these results), and then post-process (filter) the results on the client side. RGMA and the ActOn-based information service have better query abilities.

RGMA is not able to relate information available in different tables. Table 3 shows that RGMA has bad precision in query 4. It contains information to solve this query, but it comes from two different tables (`GlueCE` and `GlueSubClusterSoftwareRunTimeEnvironment`), and the query language used by RGMA does not allow joining both tables. Hence the situation is similar to the previous case: this problem can be solve on the client side by post-processing the results that have been obtained from each separate query.

RGMA is very sensitive to the registering and availability of information providers at a given point in time. Table 3 shows that RGMA has bad recall in query 1. This is because the amount of Computing Element producers that is available during the experiment is not always stable, due to the fact that either producers were not registered in the RGMA registry at that specific moment, or that the producers were not configured correctly or available at that point in time. BDII and the ActOn-based information service are more robust to this, due to the fact that they store information locally and do not depend on their information providers at the time of querying.

Some complex queries cannot be answered by one information service in isolation. Tables 2 and 3 show that BDII and RGMA can only answer the first four queries. They cannot answer queries 5 and 6 because their information providers cannot provide enough information and should be combined. This shows that the ability of BDII and RGMA to share their data resources is weak. On the other hand, the ActOn-based information service has the ability to adopt existing information sources as its information providers, and aggregate information from these information sources to answer such complex queries.

5. Lessons learned

We have gathered valuable lessons from our experience in designing the experiments for information quality measurement and conducting them on the EGEE Grid testbed. Most of them are related to the fairness of the information quality measurement process.

First, **there are not standard domain-independent methods to measure information quality in information systems.** To design an experiment in a specific domain (e.g., Grid information services), we must design it according to that domain and the information needs of the information service users.

Second, **different information services use different information models, and usually provide different expressivity in their query languages or access APIs.** Hence a special effort has to be made in order to define clearly a fair way to perform measurements that takes into account these differences.

Acknowledgements

This work is supported by the EU FP6 OntoGrid project (FP6-511513), by the Marie Curie fellowship RSSGRID (FP6-2002-Mobility-5-006668), and by the EU FP6 CoreGrid Network of Excellence (FP6-004265). We also thank Pinar Alper (IMG group), Antun Balaz and Laurence Field (EGEE porject), Georges Da Costa and Anastasios Gounaris (CoreGrid WP2), for their comments.

References

[1] X. Zhang and J. Schopf, *Performance analysis of the globus toolkit monitoring and discovery service, mds2*, in the International Workshop on Middleware Performance (MP 2004), part of the 23rd International Performance Computing and Communications Workshop (IPCCC), April 2004.

[2] *Berkeley Database Information Index (BDII)*, http://lfield.home.cern.ch/lfield/cgi-bin/wiki.cgi?area=bdiipage=documentation.

[3] E. W. Team, *EDG RGMA*, www.marianne.in2p3.fr/datagrid/documentation/rgma-guide.pdf.

[4] W. Xing, O. Corcho, C. Goble, and M. Dikaiakos, *A Grid Information Service based on an Intelligent Information Integration Architecture*, in Europe Semantic Web Conference 2007 (ESWC-2007), 2007, Poster.

[5] M. Parkin, S. van den Burghe, O. Corcho, D. Snelling, and J. Brooke, *The Knowledge of the Grid: A Grid Ontology*, in Proceedings of the 6th Cracow Grid Workshop, Cracow, Poland, October 2006.

[6] *OntoGrid CVS*, http://www.ontogrid.net/ontogrid/downloads.jsp.

[7] R. Wang and D. Strong, *Beyond Accuracy: What Data Quality Means to Data Con- sumers*, Management Information Systems, vol. 12, no. 4, pp. 534, 1996.

[8] B. Hughes, *Metadata quality evaluation: Experience from the open language archives community*, in ICADL, 2004, pp. 320329.