GLUE 2 deployment: Ensuring quality in the EGI/WLCG information system

Stephen Burke\textsuperscript{1,3}, Maria Alandes Pradillo\textsuperscript{2}, Laurence Field\textsuperscript{2} and Oliver Keeble\textsuperscript{2}

\textsuperscript{1} egi.eu, 140 Science Park, 1098 XG Amsterdam, The Netherlands
\textsuperscript{2} CERN, CH-1211 Genève 23, Switzerland

E-mail: stephen.burke@stfc.ac.uk

Abstract. The GLUE 2 information model is now fully supported in the production EGI/WLCG information system. However, to make it usable and allow clients to rely on the published information it is important that the meaning is clearly defined, and that information providers and site configurations are validated to ensure as far as possible that what they publish is correct. In this paper we describe the definition of a detailed schema usage profile, the implementation of a software tool to validate published information according to the profile and the use of the tool in the production Grid, and also summarise the overall state of GLUE 2 deployment.

1. Introduction
Version 1 of the GLUE information model has been in use since 2002, with the current version being GLUE 1.3 defined in 2007 [1]. The process to define a major new version 2.0 [2] during 2008/9 in the context of an Open Grid Forum (OGF) working group [3] was reported at CHEP ’09 [4]. The new model is not backward-compatible with the old one, and hence its introduction in a production Grid had to be carefully managed to avoid disruption to existing clients. The details of this process were reported at CHEP 2010 [5]. This paper describes the progress towards having GLUE 2 as a full production service, and in particular the steps taken to ensure the quality of the published information.

2. GLUE 2 deployment
GLUE 2 itself is an abstract model; to be used it needs to be implemented in a particular technology. For the EGI/WLCG [6] information system this consists of a set of LDAP servers known as the BDII [7]. The implementation of the LDAP schema and its deployment with the BDII in the production Grid followed rapidly after the LDAP representation of GLUE 2 [8] was defined, and was deployed in production by October 2010, although it took until late 2012 before all sites had upgraded to the new versions.

Without information providers to publish information in the GLUE 2 format the information system would be empty, and it was therefore necessary to upgrade all providers to publish in both GLUE 1 and GLUE 2 formats. These were progressively upgraded from 2010, and the process received a substantial boost from the EMI middleware project [9] which made GLUE 2 publication a major objective for the EMI 2 release in 2012. Publication has also been developed for Globus services by the IGE project [10].

\textsuperscript{3} To whom any correspondence should be addressed.
3. The EGI profile for GLUE 2
By early 2012 the deployment of GLUE 2 was reaching maturity, and EGI was considering the steps needed to make it fully usable. One aspect of this was the definition of a profile to supplement the GLUE 2 specification with more detailed information related to the specific needs of the project. A draft document was written, and after an extensive public comment period the final version was published in March 2013 [11] (although further revisions in the light of experience are possible). The following sections give more detail about the profile.

3.1. Use cases
There are several potential uses for the information represented by the GLUE 2 model, and the likely uses have implications for publication: for example, on the required accuracy, the rate at which information should be updated, the acceptable latency, and whether it may be useful to cache information if updates are temporarily blocked. There may also be implications for the future evolution of the information system, because there are potential advantages in separating different kinds of information. The profile therefore classifies the potential uses for each attribute according to the following categories:

- **Service Discovery.** These are attributes which are either likely to be used in queries which select services or metadata about the services returned by such queries, for example service endpoint types and URLs. These will generally be slowly-changing, but must be accurate as service discovery is the principle use case for the information system.

- **Service Selection.** These are attributes which reflect the current state of the service, for example the number of queued jobs in a computing system, which may be used to choose between services which are otherwise equivalent. Such information will often be rapidly-changing, and hence to be useful it needs to be fresh and reasonably accurate – old or cached values may be worse than useless.

- **Monitoring.** Such information is useful to monitor the overall state of the Grid, for example the number of running jobs in computing systems. This needs to be reasonably fresh and accurate, but is less critical than information used for service discovery.

- **Oversight.** These are a fairly small number of attributes which can be aggregated to give a high-level summary of the available resources in the Grid useful for managers, for example the installed computing and storage capacity and the service versions in use. These are usually slowly-changing and it may well be useful to cache the information. They also need to be reasonably accurate – for example, a single erroneously large value may make an aggregated sum meaningless.

- **Diagnostic.** Some information is useful to diagnose problems within the information system itself, for example creation timestamps. This is generally not a vital use case, but if the information is present it needs to be accurate.

3.2. Importance
The GLUE 2 specification marks attributes as mandatory or optional. However, in the EGI context it is useful to give more detailed guidance about the circumstances under which optional attributes may or may not be published, and the profile document therefore specifies five categories:

- **Mandatory.** This includes all attributes which are mandatory in GLUE 2 itself, but also adds some attributes which EGI regards as essential.

- **Recommended.** These are attributes which are highly desirable, but where there may be technical reasons why publication may be impossible or meaningless in particular cases.

- **Desirable.** Such information is optional, but EGI encourages publication where possible as there is a clear use for it.
• **Optional.** This is information for which there is no clear use in EGI, or where such uses would only apply in specialised situations.
• **Undesirable.** Some attributes could be actively damaging in the EGI context, either because they would lead to excessive data volumes or because they could reveal potentially sensitive information.

3.3. **Validation**

For information to be useful it needs to be accurate and meaningful, and it is therefore desirable to have criteria which can be used to validate it as far as possible. For each attribute the profile document defines tests which may be used to assess its validity. In most cases it is not possible to make the tests completely clear-cut, and they are therefore classified into four categories which may need to be treated differently:

• **FATAL.** In a few cases an invalid value may make the information completely unusable, for example unique IDs which are not in fact unique. Such cases should ideally be detected early and removed completely from the system, or otherwise need to be fixed as a matter of urgency.

• **ERROR.** This applies to cases where the value is definitely wrong, e.g. out of an allowed range, and hence must be corrected.

• **WARNING.** Values failing these tests will typically be erroneous, but there may be cases where they are in fact correct and it must therefore be possible to ignore the warning.

• **INFO.** For many attributes it is not possible to clearly define an allowed range, but nevertheless there may be heuristics which enable anomalies to be identified for further examination; for example, if an attribute representing the name of some piece of software is published by only one Grid site it may be due to a typing error. Such tests involve a trade-off between identifying as many mistakes as possible and minimising the number of false positives, and may therefore need to be tuned in the light of experience.

4. **The glue-validator tool**

The profile document specifies a large number of validation tests, and it is therefore desirable to have an automated way to check compliance. This is implemented by the glue-validator tool [12].

The glue-validator is useful for system administrators and middleware developers who want to validate whether the information published by the service they are managing or developing is compliant with the GLUE specification. It is also useful for EGI as a whole in acceptance testing for new middleware [13], and for integration into the operational monitoring of Grid sites [14]. The glue-validator is able to validate against both the basic information model definitions (for GLUE 1 and 2) and the EGI profile for GLUE 2; the latter is the recommended and default configuration.

The validator is distributed in the EMI and EGI software repositories and it will be deployed in the EGI monitoring infrastructure by November 2013. This will allow for an automated approach to full-scale GLUE 2 validation of the EGI production infrastructure.

The glue-validator is written in python and contains several libraries for each schema version, describing the schema itself (e.g. which attributes are mandatory), the existing types and the set of associated tests. The tool can be run excluding those tests which fail due to identified bugs in the information providers. This offers two different approaches to validation: system administrators can focus on fixing errors in their site configuration, ignoring those errors that are coming from the middleware itself; whereas middleware developers can run the full set of tests, identifying which aspects of the publication from their software are non-compliant.

A number of bugs in the middleware have already been identified and reported to the developers and a list of known issues is maintained [15]. Documentation has been created to give guidance to system administrators on what to do to fix the errors affecting their site [16]. The list of known issues and documentation is expected to evolve in the coming months as middleware developers fix the known issues.
5. GLUE 2 deployment status
Since early 2013 all EGI sites (including those using the NorduGrid [17] middleware) have been publishing information in the GLUE 2 format in addition to GLUE 1. Nearly all services used in the Grid are either publishing already or have publication in development. However, OSG [18] sites are currently not publishing GLUE 2 and therefore are not benefiting from the GLUE validation efforts which are only focused on GLUE 2. At the time of writing the information system contains GLUE 2 information from 2084 services at 328 sites, represented by a total of 1.8 million attributes in 191000 separate objects.

The glue-validator tool has been available since 2012. However, active publicity for the tool among sites did not begin until 2013, when tutorials and presentations have been organized. Regular automated monitoring of sites has been done since June 2013, but in a development mode without tickets being opened, and hence most sites have not been actively following the results or fixing problems. This is expected to change in early 2014 when tickets will be opened automatically from the monitoring system. Nevertheless, continuous monitoring and manual follow-up with the sites have already proven to be very useful in reducing the number of errors and improving the quality of the information.

The total number of anomalies detected by the glue-validator is currently 15253 ERRORs, 129899 WARNINGs and 85668 INFOs; if known middleware bugs are excluded there are 15187 ERRORs, 3311 WARNINGs and 21132 INFOs. However, it should be borne in mind that in some cases many separate anomalies may arise from the same underlying cause. It is expected that the numbers will reduce substantially as bugs are fixed and as the operational tests begin to raise alarms, although for INFO tests in particular some are likely to remain as the tests may be generating false positives. The tests will be adjusted as necessary if too many false positives are reported.

Most of the observed errors affect computing services, and are often related to misconfigurations in the site and in particular its underlying batch systems. The information providers rely on certain parameters that need to be configured by the site. Not all the batch systems can currently be easily configured with an automated tool and hence manual configurations are applied. These manual configurations are difficult to reproduce after an upgrade and may miss some important configuration parameters. In some cases sites use automated configuration tools incorrectly, which results in incomplete configurations. Automating the discovery of these anomalies using glue-validator should help in fixing the root causes, by producing better configuration documentation or improving the existing configuration tools.

There are also some bugs in the interpretation of batch system parameters by the information providers, which are reported to the developers as found and are usually fixed as part of the normal software improvement cycle. In addition, some elements of the publication by storage services, particularly in the context of the precise definition of attributes related to used and free space, are currently under discussion to harmonise the treatment by different storage implementations.

6. Conclusions
After many years of development, GLUE 2 is now in production use in the EGI/WLCG Grid. The EGI profile document and the glue-validator tool mean that we have a much more precise definition of what the information system should contain than we had for GLUE 1, and that the published information is validated to a much higher standard, allowing users to have confidence in the information it provides. It is expected that GLUE 2 will become the default information model in EGI during 2014, completing a process which started in 2007.

Acknowledgments
This work is supported by the EGI-InSPIRE (Integrated Sustainable Pan-European Infrastructure for Researchers in Europe) project co-funded by the European Commission, contract number: RI-261323.
References
[1] Andreozzi S, Burke S, Donno F, Field L, Fisher S, Jensen J, Konya B, Litmaath M, Mambelli M, Schopf J, Viljoen M, Wilson A and Zappi R 2007 GLUE Schema Specification - Version 1.3 - draft 3
http://forge.cnaf.infn.it/plugins/scmsvn/viewcvs.php/*checkout*/v_1_3/spec/pdf/GLUESchema.pdf?rev=49&root=glueschema
[2] Andreozzi S, Burke S, Field L, Galang G, Konya B, Litmaath M, Millar P and Navarro J P 2009 GLUE Specification v. 2.0 OGF Proposed Recommendation GFD 147
http://www.ogf.org/documents/GFD.147.pdf
[3] GLUE OGF working group http://forge.ogf.org/sf/projects/glue-wg
[4] Burke S, Donno F, Ehm F, Field L, Litmaath M and Millar P 2010 The impact and adoption of GLUE 2.0 in the LCG/EGEE Production Grid J. Phys.: Conf. Ser. 219 062005
[5] Burke S, Field L and Horat D 2011 Migration to the GLUE 2.0 information schema in the LCG/EGEE/EGI production Grid J. Phys.: Conf. Ser. 331 062004
[6] EGI (European Grid Infrastructure) http://www.egi.eu/
WLCG (Worldwide LHC Computing Grid) http://www.cern.ch/lcg
[7] BDII http://gridinfo.web.cern.ch/
[8] Andreozzi S, Burke S, Ehm F, Field L, Galang G, Horat D, Konya B, Litmaath M, Millar P and Navarro J P 2010 GLUE v. 2.0 – Reference Implementation of an LDAP Schema OGF Draft Recommendation
http://forge.ogf.org/sf/go/doc15518?nav=1
[9] EMI (European Middleware Initiative) http://www.eu-emi.eu/
[10] IGE (Initiative for Globus in Europe) http://www.ige-project.eu/home
[11] Burke S 2013 EGI profile for the use of the GLUE 2.0 Information Schema http://go.egi.eu/glue2-profile
[12] glue-validator http://gridinfo.web.cern.ch/glue/glue-validator-guide
[13] EGI Quality Criteria for verification of software products http://egi-qc.github.io/
[14] EGI middleware monitoring tests https://wiki.egi.eu/wiki/MW_SAM_tests
[15] GLUE 2 publication known issues https://twiki.cern.ch/twiki/bin/view/EGEE/ISproviders#GLUE_2_0_Validation
[16] glue-validator error codes https://twiki.cern.ch/twiki/bin/view/EGEE/GLUEValidatorErrorCodes
[17] NorduGrid Collaboration http://www.nordugrid.org/
[18] OSG (Open Science Grid) http://www.opensciencegrid.org/