Report on standards developments

Project Number:	FP6-2005-IST-026996
Deliverable id:	D 3.3
Deliverable name:	Report on standards developments
Submission Date:	31/03/2008

COVER AND CONTROL PAGE OF DOCUMENT	
Project Acronym:	ACGT
Project Full Name:	Advancing Clinico-Genomic Clinical Trials on Cancer: Open Grid Services for improving Medical Knowledge Discovery
Document id:	D 3.3
Document name:	Report on standards developments
Document type (PU, INT, RE)	INT
Version:	1
Submission date:	03/03/2008
Editor: Organisation: Email:	Jarek Nabrzyski PSNC naber@man.poznan.pl

Document type PU = public, INT = internal, RE = restricted

ABSTRACT

This deliverable presents involvement of ACGT consortium members in development of standards.

KEYWORD LIST:Grid

MODIFICATION CONTROL			
Version	Date	Status	Author
1.0	20/03/2008	Draft	J. Nabrzyski
2.0	30/03/2008	Draft	J. Nabrzyski

List of Contributors

- Jarek Nabrzyski, PSNC
- Erwin Bonsma, Philips
- Stelios Sfakianakis, FORTH
- Albeto Anguita, UMA
- Juliusz Pukacki, PSNC

Contents

Executive Summary

ACGT is an Integrated Project (IP) funded in the 6th Framework Program of the European Commission under the Action Line "Integrated biomedical information for better health". The high level objective of the Action Line is the development of methods and systems for improved medical knowledge discovery and understanding through integration of biomedical information (e.g. using modelling, visualization, data mining and grid technologies). Biomedical data and information to be considered include not only clinical information relating to tissues, organs or personal health-related information but also information at the level of molecules and cells, such as that acquired from genomics and proteomics research.

ACGT focuses on the domain of Cancer research, and its ultimate objective is the design, development and validation of an integrated Grid enabled technological platform in support of post-genomic, multi-centric Clinical Trials on Cancer. The driving motivation behind the project is our committed belief that the breadth and depth of information already available in the research community at large, present an enormous opportunity for improving our ability to reduce mortality from cancer, improve therapies and meet the demanding individualization of care needs.

1. Biomedical Standards

Using standards where possible is important, especially in medical world. Data integration and tool compatibility is a serious issue here. This is why tha CGT project is so much focused on standardization activities. Within the project it is PSNC which is responsible for standards monitoring and reporting. The main problem is that today there is no unifying infrastructure or common standards for the technologies that most cancer researchers use. This means that researchers cannot share their data or benefit from the innovative informatics tools that are been developed by other researchers

Many standards or de-facto standards are being used and/or promoted by ACGT. Relevant Standards to which ACGT plans to contribute include HL7, specifically the new HL7 Clinical-Genomics Genotype model, electronic health records (i.e. extensions to RIM so that genomic data is supported in future Electronic Health Record implementation), models to represent genomic information, such as the new PML (the Polymorphism Markup Language, developed by a consortium of institutions over the world), knowledge representation standards for guidelines and protocols (e.g., GLIF, the Guideline interchange format), health professional and patient intelligent cards, network protocols other protocols. For this issue, ontologies and Grid, as expanded below, will be also key technologies to consider, enhance and adapt.

FORTH, Fhg, INRIA and other ACGT partners are already contributing and voting members of HL7 and its international affiliates groups, and will coordinate liaising with its relevant technical working groups.

2. Grid Standards

Key to the realization of the Grid vision is standardization, so that the diverse components that make up a modern computing environment can be discovered, accessed, allocated, monitored, accounted for, billed for, etc., and in general managed as a single virtual system —even when provided by different vendors and/or operated by different organizations. Standardization is critical if we are to create interoperable, portable, and reusable components and systems; it can also contribute to the development of secure, robust, and scalable Grid systems by facilitating the use of good practices.

Grid (Services) Computing is based on an open set of standards and protocols (i.e., Open Grid Services Architecture: OGSA) that enable communication across heterogeneous, geographically dispersed IT environments. The current trend is to produce a broader set of standards that cover all aspects of Grid technologies (computational, data storage, networking and web services). This effort is articulated through the Global Grid Forum.

ACGT will get involved in the Global Grid Forum that assembles grid researchers from Europe, the US and Japan. The project intends to work together with initiatives such as the Global Grid Forum's Semantic Grid Research Group [1]. The proposed work requires expertise that is hard to find in a single European country in that it will be setting up standards and protocols that will require a pan-European and global adoption, in order to make significant impact. All consortium members are active in other national and international research projects (see section B5) and will use those contacts to promote the project and its findings.

PSNC is co-founder of the Global Grid Forum and a member of its staff was a member of the Global Grid Forum Steering Group. ACGT will access to these forums to promote the international relevance of ACGT and contribute to the emerging standards.

The focus of the standardization contribution of ACGT will be the proposal of semantic extensions to the OGSA specification. Currently OGSA's ontology is technical - it speaks of services, protocols, processes, computers etc. We propose to build the semantic depth into the core of the grid standards. Our current idea is to allow for an arbitrary ontology, specified in one of the well established ontology languages, become part of the very fabric of the grid.

Specifically, ACGT participates actively in the following working/research groups of the Global Grid Forum:

- Grid Scheduling Ontology Working Group (proposed),
- Semantic Grid Research Group,
- Life Sciences Research Group
- JSDL
- GSA

The roles of these groups are to produce ontology of Grid accompanied by a set of documents describing the ontology and the tools/libraries used to create the ontology and to make use of the ontology later. Using the ontology generated by the working groups when designing and implementing the next generation of Resource Management Systems and their corresponding Grid services may further lead to ontology-driven systems.

The goal of the Semantic Grid Research Group (SEM-GRD) is to realise the added value of Semantic Web technologies for Grid users and developers. It provides a forum to track Semantic Web community activities and advise the Grid community on the application of Semantic Web technologies in Grid applications and infrastructure, to identify case studies and share good practice.

ACGT involvement (official roles) in standardization groups:

Partner	Person	Body	Group	Contribution	Specification
PSNC	Jarek Nabrzyski	OGF	Grid Scheduling Architecture Research Group (GSA-RG)	Co-founder, Ex-chair	
PSNC	Ariel Oleksiak	OGF	Grid Scheduling Architecture Research Group (GSA-RG)	Secretary	GSA Requirements (ongoing)
PSNC	Michael Russell	OGF	Grid Computing Environments	Contributor	GridSphere Portal
PSNC	Ariel Oleksiak	OGF	JSDL	Contributor	JSDL specification
PSNC	Krzysztof Kurowski, Piotr Domagalski	OGF	DRMAA	Contributor	DRMAA specification

Other standardization activities

- PSNC participates in OGF-Europe, the initiative whose main objective is pervasive Grid adoption through interoperable software standards. To realize this goal OGF-Europe organizes outreach seminars and workshops, adoption challenges and recommendations reports, community surveys, and best practice reports and tutorials. PSNC is responsible for student and expert scholarships, which are intended to increase European involvement in OGF and, on the other hand, adoption of Grid standards in Europe. PSNC is currently organizing OGF23 scholarships. To this end, PSNC met OGF authorities at OGF22.

- PSNC is an early adopter of JSDL and OGSA BES specifications. Based on implementation experiences it provided a number of comments concerning these specifications to the JSDL and OGSA BES mailing lists.

- PSNC is also, as a standardization leader in the EU BREIN and EU ACGT projects, member of Grid Standards Coordination Group (GSCG), which aims to coordinate standardization between FP6 IST Grids Unit Projects.

Participation in OGFs:

Body	Event	Partner	Person	Group	Contribution	Notes
Open Grid Forum	OGF20	PSNC	Ariel Oleksiak	GSA-RG	Presentation Secretary	Currently the group is working on interoperability of multiple Grid schedulers. One of aspects, I'm focused on, is a definition of scheduling attributes.
				GRAAP-WG	Participation	Several implementations of WS-Agreement were presented. There are issues concerning interoperability and lack of some standardized terms. Negotiation is still discussed. It seems that multiple negotiation schemes rather than one fixed should be available.
				Dynamic SLA Negotiations	Participation	Presentation of approaches for dynamic SLA negotiations
				JSDL-WG	Participation	Among others JSDL extensions discussed.
				OGSA-BES-WG	Participation	It seems that this specification has already many implementations.
				WFM-RG	Participation	The topic was sharing workflows. The group meets quite rarely, however may be related to things we are going to do within BREIN.
Open Grid Forum	OGF21	PSNC	Jarek Nabrzyski	GSA, OGSA, BSE	Participation	
Open Grid Forum	OGF22	PSNC	Ariel Oleksiak	GSA	Secretary / Preparation	Discussion about OGF specifications needed to provide interoperability between grid schedulers. Several scenarios for interoperability test discussed.
				JSDL / Activity Instance BoF	Participation	Discussions of parameter sweep extension to JSDL. Activity Instance BoF presenting proposal of a model and language to describe the whole lifecycle of an activity.
				OGSA-RSS	Participation	Results from public comment period; plans for how to split spec up into manageable pieces and which parts to pursue first. Plan to get some parts back to public comment in this year (2008)
				HPC Profile	Participation	Some additional features proposed such as application templates, advanced filter profile, data staging.

3. Best practices and standards

The ACGT project promotes the standardization of Nephroblastoma studies in Europe which at the moment suffer from difficulties rising from the integration of heterogeneous info from geographically distributed sites. ACGT develops an integrated clinico-genomic environment that will allow the clinician to access both clinical and genomic data related to nephroblastoma. The project contributes to the standardization of clinico-genomic information through the use of common vocabularies and ontologies and provide a virtual environment that will link data and expertise related to nephroblastoma from different sites. Within ACGT a database for the nephroblastoma trial and study is being specified and implemented, fulfilling Good Clinical Practices (GCP), RDE (remote data entry), security and anonymization criteria. Crucially, the ACGT environment will also encourage scientists to integrate molecular data into existing (or new) databases, and will provide an easy way to combine it with clinical data. This will be a novelty for nephroblastoma clinical trials and will enhance current clinical studies so as to correlate molecular findings with clinical data.

These novel procedures will serve as model for other clinical trials and will prove the flexibility of the ACGT environment. The nephroblastoma database and CRO will be validated within the SIOP 2001/GPOH trial in Germany and afterwards be extended to other studies, trials and clinics. The ultimate goal will be to expand it throughout the whole of Europe (to other SIOP trials in Europe) aiming that it will become a European standard for nephroblastoma trials and a model for similar trials related to cancer.

References

[1] http://www.semanticgrid.org/GGF/

Appendix A - Abbreviations and acronyms

SOA	Service Oriented Architecture
JSDL	Job Submission Description Language
GCP	Good Clinical Practices
GSA	Grid Scheduling Architecture
OGF	Open Grid Forum
DRMAA	Distributed Resource Management Architecture API
WSDL	Web Service Definition Language
OGSA	Open Grid Services

APPENDIX B Standards used in ACGT

1. WSDL

Web Services Description Language - XML based language for describing interfaces of Web services. The WSDL defines services as collections of network endpoints, or ports. The WSDL specification provides an XML format for documents for this purpose. The abstract definition of ports and messages is separated from their concrete use or instance, allowing the reuse of these definitions. A port is defined by associating a network address with a reusable binding, and a collection of ports define a service. Messages are abstract descriptions of the data being exchanged, and port types are abstract collections of supported operations. The concrete protocol and data format specifications for a particular port type constitutes a reusable binding, where the messages and operations are then bound to a concrete network protocol and message format. In this way, WSDL describes the public interface to the web service.

WSDL is the most important standard used for Web services implementation. In ACGT project all services implemented are described using WSDL.

2. GSI

Grid Security Infrastructure is a specification for secret, tamper-proof, delegatable communication between software in a grid computing environment. Secure, authenticatable communication is enabled using asymmetric encryption.

A central concept in GSI authentication is the certificate. Every user and service on the Grid is identified via a certificate, which contains information vital to identifying and authenticating the user or service. A GSI certificate includes four primary pieces of information:

- subject name, which identifies the person or object that the certificate represents.

- the public key belonging to the subject.

- the identity of a Certificate Authority (CA) that has signed the certificate to certify that the public key and the identity both belong to the subject.

- the digital signature of the named CA.

The GSI uses the Secure Sockets Layer (SSL) for its mutual authentication protocol. By default, the GSI does not establish confidential (encrypted) communication between parties. Once mutual authentication is performed, the GSI gets out of the way so that communication can occur without the overhead of constant encryption and decryption.

The GSI provides a delegation capability: an extension of the standard SSL protocol which reduces the number of times the user must enter his pass phrase. If a Grid computation requires that several Grid resources be used (each requiring mutual authentication), or if there is a need to have agents (local or remote) requesting services on behalf of a user, the need to re-enter the user's pass phrase can be avoided by creating a proxy.

In ACGT project GSI was introduced as a common security infrastructure not only for the services of Grid layer, but for all services in ACGT environment.

3. JSDL

Job Submission Description Language is an extensible XML specification from the Global Grid Forum for the description of simple tasks to non-interactive computer execution systems. Currently at version 1.0 (released November 7, 2005), the specification focuses on the description of computational task submissions to traditional high-performance computer systems like batch schedulers.

JSDL describes the submission aspects of a job, and does not attempt to describe the state of running or historic jobs. Instead, JSDL includes descriptions of:

- Job name, description
- Resource requirements that computers must have to be eligible for scheduling, such as total RAM available, total swap available, CPU clock speed, number of CPUs, Operating System, etc.
- Execution limits, such as the maximum amount of CPU time, wallclock time, or memory that can be consumed.
- File staging, or the transferring of files before or after execution.
- Command to execute, including its command-line arguments, environment variables to define, stdin/stdout/stderr redirection, etc.

In ACGT project JSDL is used by resource management system for Grid (GRMS) that is used for submission of computational jobs to the Grid.

4. SPARQL

Within ACGT we use SPARQL [1] as a query language. It is supported by the data access services that wrap the various data sources. The semantic mediator uses SPARQL as well, not only in constructing the queries that are sent to the data access service, but also for receiving queries (expressed in the ACGT Master Ontology).

The choice for SPARQL as the common query language was made after first identifying the main requirements, and subsequently evaluating various candidate query languages, including SQL and XQuery. SPARQL was judged the most suitable, as it is based on a general graph-based model, and because it is less complex than the other query languages. SPARQL has an intermediate level of expressiveness, which was also deemed appropriate given the range of data sources that need to be integrated into the ACGT architecture. A more detailed justification for the use of SPARQL can be found in D5.2 [2], which also gives examples of how it is used.

For returning the results of SPARQL queries we are using the SPARQL Query Results XML Format [3], which is as the name suggests a way to return the query results using XML. Given the choice to use Web Services and SPARQL as the common query language, this is therefore a logical choice. D5.2 includes an example that shows how results are formatted using the SPARQL Query Results XML Format.

5. RDFS

RDFS is a knowledge representation language, which final W3C recommendation was released in February 2004. It is designed to describe the schema of an RDF repository. RDF is a general-purpose language that allows describing resources. Its XML syntax is designed to permit a high compatibility among different applications, even when covering heterogeneous domains. RDF has been developed by the RDF Core Working Group as part of the W3C Semantic Web Activity. In RDF, resources are divided in classes. A class defines a group of elements sharing the same properties. Each class may have a series of attributes called *properties*. These properties can connect a class with another class or with a basic type. Classes are arranged hierarchically, so that subclasses of a given class inherit the properties of that class.

The OWL language for describing ontologies is built on top of RDFS. It inherits its syntax, adding some extra capability for defining constraints and different types of properties. This language was chosen in ACGT to specify the Master Ontology. OWL is currently the most used ontology language.

RDFS defines a series of XML tags that allow specifying the classes and properties that an RDF document can make use of. Next, a brief description of the most important tags is presented:

1. rdfs:Class: allows defining RDF classes. The name of the class is specified though attributes inside that tag
2. rdfs:subClassOf: this tag is placed inside a class declaration, and allows defining the class hierarchy by indicating the superclass of the class being defined
3. rdf:Property: allows defining an RDF Property. Detailed information, such as domain and range, is specified by means of other tags
4. rdfs:domain: declares the domain of a property (usually a class)
5. rdfs:range: declares the range of a property

In ACGT, RDFS is used to describe the schemas of the databases to be integrated in the Semantic Mediator. In some way, it is also utilized to describe the ACGT Master Ontology—in fact, OWL-DL is the language in which the MO is written, however this is not but an extension of the RDFS language, as it was explained before.

RDFS is also the language for exposing the Global Schema—a schema representing the set of queries accepted by the mediator. This schema is generated automatically from the mappings of the integrated databases with the MO. To create it, two steps are carried out sequentially: i) the RDFS information of the MO is extracted, generating an RDFS document —unnecessary information for the global schema, such as some types of OWL-DL specific restrictions, is eliminated— and ii) intersection of this information with the set of mappings with databases is generated, providing the final Global Schema to be used by the Semantic Mediator.

6. BPEL

In the context of WP9 'The Integrated ACGT Environment' the Web Services Business Process Execution Language (WS-BPEL) is used. Standardized by the OASIS organization, WS-BPEL is a language for specifying business process behaviour based on Web Services. It defines an interoperable integration model that should facilitate the expansion of automated process integration in both the intra-corporate and the business-to-business spaces. The processes described in WS-BPEL can be one of two kinds: Executable and Abstract processes. Executable business processes model actual behaviour of a participant in a business interaction. Abstract business processes are partially specified processes that are not intended to be executed and they may be used to hide some of the required concrete operational details.

In the ACGT platform WS-BPEL is the technology used to support the high level integration of the ACGT services and tools into complex scientific workflows for the implementation, testing and validation of user scenarios. The workflows defined in WS-BPEL are deployed as executable processes and they can be subsequently used as atomic services to construct even more complex and higher level scenarios.

Much of the effort in ACGT that is related to WS-BPEL has been focused on making the tools that implement this standard more compliant with the rest of the ACGT architecture. Notable work areas are the invocation of Grid Services and the integration of Grid Security Infrastructure.