Building a European Biomedical Grid on Cancer: The ACGT Integrated Project

M. Tsiknakis[1], D. Kafetzopoulos[2], G. Potamias[3], A. Analyti[4], K. Marias[3], A. Manganas[3]

[1]Center for eHealth Technologies, Institute of Computer Science, FORTH, Crete, Greece

[2] Post Genomic Technologies Laboratory, Institute of Molecular Biology and Biotechnology, FORTH, Crete, Greece

[3] Biomedical Informatics Laboratory, Institute of Computer Science, FORTH, Crete, Greece

[4] Information Systems Laboratory, Institute of Computer Science, FORTH, Crete, Greece

ABSTRACT

This paper presents the needs and requirements that led to the formation of the *ACGT* (Advancing Clinico Genomic Trials) integrated project, its vision and methodological approaches of the project. The ultimate objective of the *ACGT* project is the development of a European biomedical grid for cancer research, based on the principles of open access and open source, enhanced by a set of interoperable tools and services which will facilitate the seamless and secure access to and analysis of multi-level clinico-genomic data, enriched with high-performing knowledge discovery operations and services.

By doing so, it is expected that the influence of genetic variation in oncogenesis will be revealed, the molecular classification of cancer and the development of individualised therapies will be promoted, and finally the in-silico tumour growth and therapy response will be realistically and reliably modelled.

Its main design decisions and results at its current stage of development are presented.

KEYWORDS

Biomedical grids, Semantic data mediation and integration, Data mining and knowledge discovery on the Grid, Cancer research

1. Introduction

This is a critical time in the history of cancer research as recent advances in methods and technologies have resulted in an explosion of information and knowledge about cancer and its treatment. As a result, our ability to characterize and understand the various forms of cancer is growing exponentially, and cancer therapy is changing dramatically. Today, the application of novel technologies from proteomics and functional genomics to the study of cancer is slowly shifting to the analysis of clinically relevant samples such as fresh biopsy specimens and fluids, as the ultimate aim of translational research is to bring basic discoveries closer to the bedside.

The implementation of *discovery driven translational research*, however, will not only require co-ordination of basic research activities, facilities and infrastructures, but also the creation of an *integrated and multidisciplinary environment* with the participation of dedicated teams of clinicians, oncologists, pathologists, epidemiologists, molecular biologists, as well as a variety of disciplines from the domain of information technology.

Today, information arising from post-genomics research, and combined genetic and clinical trials on one hand, and advances from high-performance computing and informatics on the

other is rapidly providing the medical and scientific community with new insights, answers and capabilities. The breadth and depth of information already available to the research community at large, presents an enormous opportunity for improving our ability to reduce mortality from cancer, improve therapies and meet the demanding individualization of care needs. A critical set of challenges, however, currently inhibit our capacity to capitalize on these opportunities [1]. Much of the genomic data of clinical relevance generated so far are in a format that is inappropriate for diagnostic testing. Very large epidemiological population samples followed prospectively (over a period of years) and characterized for their *biomarker* and *genetic variation* will be necessary to demonstrate the clinical usefulness of these tools.

Up to now, the lack of a common infrastructure has prevented clinical research institutions from mining and analyzing disparate data sources. This inability to share technologies and data developed by different cancer research institutions can therefore severely hamper the research process. Similarly, the lack of a unifying architecture is proving to be a major roadblock to a researcher's ability to mine different databases. Most critically, however, even within a single laboratory, researchers have difficulty integrating data from different technologies because of a lack of common standards and other technological and medico-legal and ethical issues.

As a result, very few cross-site studies and clinical trials are performed and in most cases it isn't possible to seamlessly integrate multi-level data (from the molecular to the organ, individual and population levels). In conclusion, clinicians or molecular biologists often find it hard to exploit each other's expertise due to the absence of a cooperative environment which enables the sharing of data, resources or tools for comparing results and experiments, and a uniform platform supporting the seamless integration and analysis of disease-related data at all levels.

1.1 GRID Computing

Grid computing enables the virtualization of distributed computing over a network of heterogeneous resources giving users and applications seamless, on demand access to vast IT capabilities [2]. Grid computing provides a novel approach to harnessing distributed resources, including applications, computing platforms or databases and file systems. Applying grid computing can drive significant benefits by improving information access and responsiveness, and adding flexibility, all crucial components of solving the data warehouse dilemma.

Rather than bringing data to a data warehouse where it sits waiting to be used, a federated solution can maintain the data at its points of origin. Federated solutions help to address the size and complexity of data warehouses by applying a logical model to the existing physical infrastructure instead of imposing a new data warehouse environment. Information grid technology—which gives users and applications security-rich access to virtually any information source, anywhere, over any type of network—supports sharing of data for processing and large-scale collaboration. It also helps bring the federated model to distributed and complex data sources.

Grid computing, also, introduces a new concept to IT infrastructures because it supports distributed computing over a network of heterogeneous resources and is enabled by open standards. As a result, new and innovative approaches are evolving for harnessing the vast and unused computational power of the world's computers and direct it at research designed to help unlock genetic codes that underlie diseases like cancer, AIDS and Alzheimer's.

2. A European biomedical Grid infrastructure for Clinical Trials on Cancer: The ACGT Vision

Within such a context, the implementation of the EU funded Integrated Project named "*Advancing Clinico-Genomic Trials on Cancer: Open Grid Services for Improving Medical Knowledge Discovery*", with the acronym *ACGT*, is beginning.

The ultimate objective of the *ACGT* project is the provision of a unified technological infrastructure which will facilitate the seamless and secure access and analysis of multi-level clinico-genomic data enriched with high-performing knowledge discovery operations and services (see Fig. 1). In so doing, *ACGT* aims to contribute to (a) the advancement of cancer research for revealing the influence of genetic variation in oncogenesis, (b) the promotion of molecular classification of cancer and the development of individualised therapies, and (c) the development of realistic and reliable in-silico tumour growth and therapy response models (for the avoidance of expensive and often dangerous examinations and trials on patients) [3].

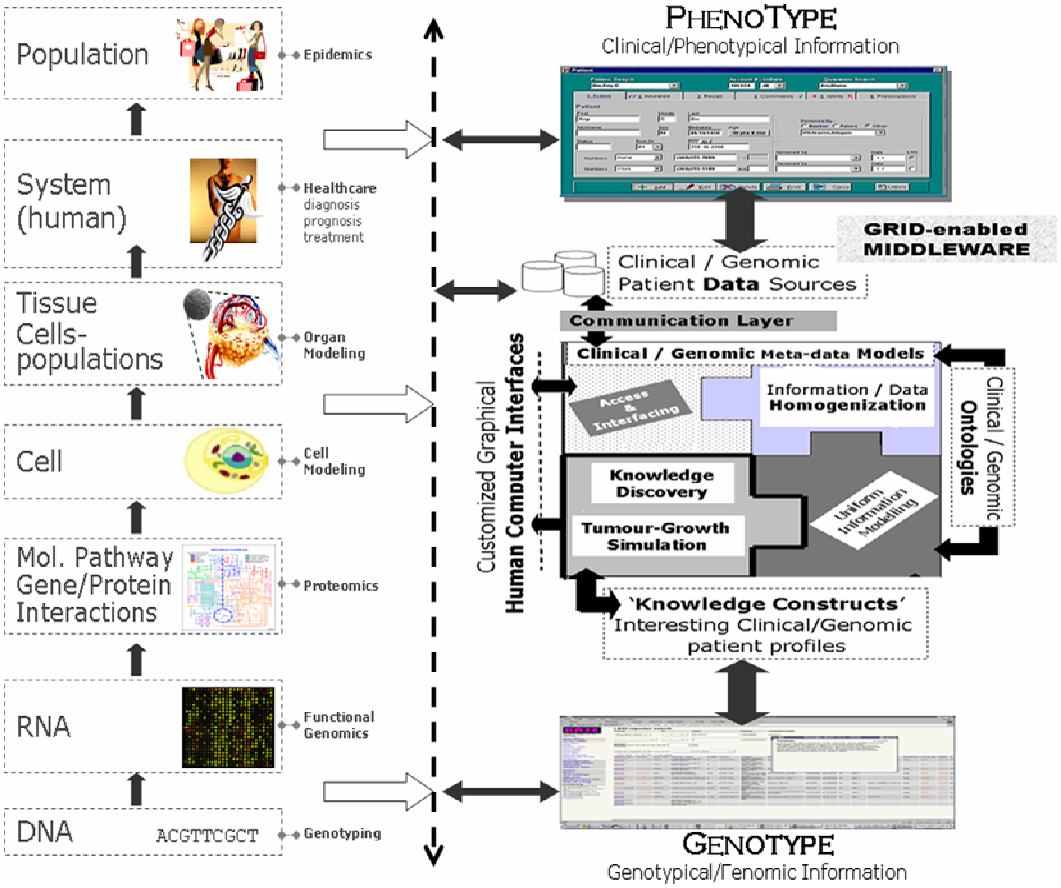


Figure 1: The envisioned ACGT GRID-enabled infrastructure and integrated environment – integration to be achieved at all levels, from the molecular to system and to the population.

The real and specific problem that underlies the *ACGT* concept is *"co-ordinated resource sharing and problem solving in dynamic, multi-institutional, Pan-European virtual organisations"*. This sharing is, necessarily, highly controlled, with resource providers and consumers defining clearly and carefully just what is shared, who is allowed to share, and the conditions under which sharing occurs. A set of individuals and/or organisations defined by such sharing form what we call the *ACGT virtual organisation (VO).*

The project intends to eventually become a Pan-European community of voluntary participants from national and trans-national biomedical research fields, given the benefits of

open access to a rich source of interoperable tools, shared data and standards developed by the BioMedical Informatics (BMI) research community.

The basic principles on which *ACGT* is basing its R&D and service delivery vision are:

- Clinical Research Organisations (CROs) will continue to retain their independence, whilst their collaboration with each other will be determined by their interests;
- The technological infrastructure deployed by the various CROs will be different, hence heterogeneous;

In achieving the above objectives, we envisage a need for:

- highly flexible and dynamic sharing relationships. The dynamic nature of sharing relationships means that we require mechanisms for discovering and characterising the nature of the relationships that exist at a particular point in time. For example, a new participant joining a VO must be able to determine what resources it is able to access, the "quality" of these resources, and the policies that govern access;
- sophisticated and precise levels of control over how shared resources are used, including fine-grained and multi-stakeholder access control, delegation, and application of local and global policies;
- sharing of varied resources, ranging from programs and data to computers;
- diverse usage models, ranging from single user to multi-user and from performance sensitive to cost-sensitive.

Consequently, ACGT will create and test an infrastructure for cancer research by using a virtual web of trusted and interconnected organizations and individuals to leverage the combined strengths of cancer centres and investigators and enable the sharing of biomedical cancer-related data and research tools in a way that the common needs of interdisciplinary research are met and tackled. Furthermore, *ACGT* intends to build upon the results of several biomedical Grid projects and initiatives, such as the caBIG [4], BIRN [5], MEDIGRID [6], MyGRID [7] and DiscoverySpace [8].

The project focuses on the semantically rich problems of dynamic resource discovery, workflow specification, and distributed query processing, as well as provenance management, change notification, and personalization. The infrastructure work in *ACGT* contains the following main components:

- **BIOMEDICAL TECHNOLOGY GRID LAYER:** This layer comprises the basic "*Grid engine*" for the scheduling and brokering of resources. This layer enables the creation of "*Virtual Organisations (VO)*" by integrating users from different and heterogeneous organisations. Access rights, security (encryption), trust buildings are issues to be addressed and solved on this layer, based on system architectural and security analysis.
- **DISTRIBUTED DATA ACCESS AND APPLICATIONS:** In order to provide seamless and interoperable data access services to the distributed data sources, a set of compatible software key modules/services will be developed based on Web Services. These services will provide ontology-based ubiquitous interoperability among the integrated *ACGT* environment and other types of heterogeneous information systems, i.e. clinical, LIMS, microarray, SNP/genotyping, etc.
- **DATA MINING AND KNOWLEDGE DISCOVERY TOOLS:** The "Data mining and Knowledge Discovery Services" layer includes open data mining and data analysis services. *ACGT*

will devote significant effort towards the design, development and deployment of *open, interoperable* data mining and analysis software tools and services. The ultimate goal is to offer a GRID-enabled *Knowledge Discovery Suite* [9] for supporting discovery operations from combined clinico-genomic biomedical data.

- **ONTOLOGIES AND SEMANTIC MEDIATION TOOLS:** Formalised knowledge representations (ontologies) will play a key role in any future biomedical Grid on cancer research. This creates the requirement for adopting/extending or even constructing an ontology for the particular disease under investigation. By building on the various ontologies and controlled vocabularies that have grown over the years for providing a shared language for the communication of biomedical information (e.g., the Gene Ontology (GO), the MGED Ontology, the NCI Thesaurus and Metathesaurus, the UMLS Metathesaurus, etc.), *ACGT* is devoting significant R&D effort to the task of constructing a shared ontology for the disease under investigation.
- **TECHNOLOGIES AND TOOLS FOR IN-SILICO ONCOLOGY:** *ACGT* will demonstrate its added value for the in-silico modelling of tumour growth and therapy response. The aim been to develop open tools and services for the four dimensional, patient specific modelling and simulation of the biological activity of malignant tumours and normal tissues in order to optimize the spatiotemporal planning of various therapeutic schemes. Ultimately, the aim of this activity is to contribute to the effective treatment of cancer and to contribute to the understanding of the disease at the *molecular, cellular, and higher level(s) of complexity*.
- **THE INTEGRATED *ACGT* ENVIRONMENT:** Integration of applications and services will require substantial meta-information on algorithms and input/output formats if tools are supposed to interoperate. Assembly of tools for virtual screening into complex workflows will only be possible if data formats are compatible and semantic relationship between objects shared or transferred in workflows are clear.

3. R&D Challenges and the ACGT approach

A major part of the project is devoted to research and development in infrastructure components that eventually will be integrated into a workable demonstration platform upon which the selected, and those to be selected during the lifecycle of the project, Clinical Trials can be demonstrated and evaluated against user requirements defined at the onset of the project.

ACGT's vision is to become a pan-European voluntary network of grid connecting individuals and institutions to enable the sharing of data and tools, creating a European Wide Web of cancer clinical research; the ultimate goal being to speed the delivery of innovative approaches for the prevention and treatment of cancer.

In realizing this vision the *ACGT* project must move beyond the current state of the art in a number of domains. Some of the challenges facing *ACGT* and the approach taken in tackling those challenges are briefly described in the following sections.

3.1 The ACGT Clinical Trials

Today it is recognised that the key to individualizing treatment for cancer lies in *translational research*, i.e. in finding ways to quickly "translate" the discoveries about human genetics made by laboratory scientists in recent years into tools that physicians can use to help make decisions about the way they treat patients.

In Europe there are several ongoing clinical studies related to cancer. However, amongst the different hospitals involved there is heterogeneity in the way patient data is documented while

electronic patient records aren't available in all hospitals. In several ongoing clinical trials "case report forms" (CRFs) are frequently used to record protocol-specified data about patients.

The use of distributed approaches introduces significant advantages that should be analysed in the context of clinical trials. The new scenarios of genomic medicine introduce significant new challenges that cannot be addressed with our current methodologies. One of the issues where the *ACGT* project presents an innovative approach is the design of new, combined clinico-genomic translational clinical trials, enabled by the set of innovative tools and services available to all members of the "virtual organisation" created through the use of GRID.

There are three main *clinico-genomic trials* (C-GT) in the *ACGT* project. The realization of these trials will be based on a number of *scenarios* which will act as *benchmark* references for the development and assessment of the *ACGT* technology. On the systems' level, these scenarios will guide the specification, the development and the evaluation of the GRID-enabled *ACGT* integrated environment and platform. On the clinical and genomics levels, these scenarios will offer clear-cut references for assessing the reliability of the *ACGT*-based clinico-genomic trials' outcome.

1. The first C-GT focuses on ***breast cancer*** (BC) and addresses the *predictive* value of gene-expression profiling (based on microarrays and genotyping technology) in classifying (according to induced 'good' and 'bad' prognostic molecular signatures) and treating breast cancer (BC) patients.
2. The second C-GT focuses on ***paediatric nephroblastoma*** or, ***Wilms tumour*** (PN) and addresses the treatment of PN patients according to well-defined risk groups in order to achieve the highest possible cure rates, to decrease the frequency and intensity of acute and late toxicity and to minimize the cost of therapy. The main objective of this trial is to explore and offer a molecular extension dimension to PN treatment harmonized with traditional clinico-histological approaches.
3. The third C-GT focuses on the development and evaluation of ***in silico*** tumour *growth* and tumour/normal tissue *response simulation* models – in silico tumour growth and simulation modelling (IS-TGSM). The aim of this trial is to develop an 'oncosimulator' and evaluate the reliability of *in-silico* modelling as a tool for assessing alternative cancer treatment strategies.

3.2 The Biomedical Grid and Heterogeneous Data Integration

We have selected the Globus Toolkit for the implementation of the grid middleware for building our open grid layer. The Globus Toolkit [10] is an open source software toolkit developed by the Globus Alliance and many others. The Globus Toolkit provides grid services that meet the requirements of the Open Grid Service Architecture and are implemented on top of the Web Service Resource Framework. It includes software for security, information infrastructure, resource management, data management, communication, fault detection, and portability.

The most important components of the Globus Toolkit involved in our envisaged grid system is WS-GRAM (Web Services – Grid Resource Allocation & Management) for job execution, MDS4 (Monitoring & Discovery System) and GSI (Grid Security Infrastructure). Other technologies that will be included in ACGT are Globus security and OGSA-DAI as a grid data layer for exposing data services. The OGSA-DAI data service is responsible for accessing and retrieving clinical and genomic information from the corresponding information systems [11].

A critical feature of *ACGT* however is to create semantic interoperability between data resources. As part of the grid architecture, the *ACGT* Master Ontology is a central part in creating the semantic interconnection. Classical approaches to database integration [12] include techniques such as wrappers or virtual conceptual schemas. Ontologies are a relevant method for database integration and, in fact, many current projects and proposals are evolving towards ontology-based methods. By using these ontology-based approaches, developers can map, for instance, objects belonging to a specific database to concepts of a shared ontology or biomedical vocabulary.

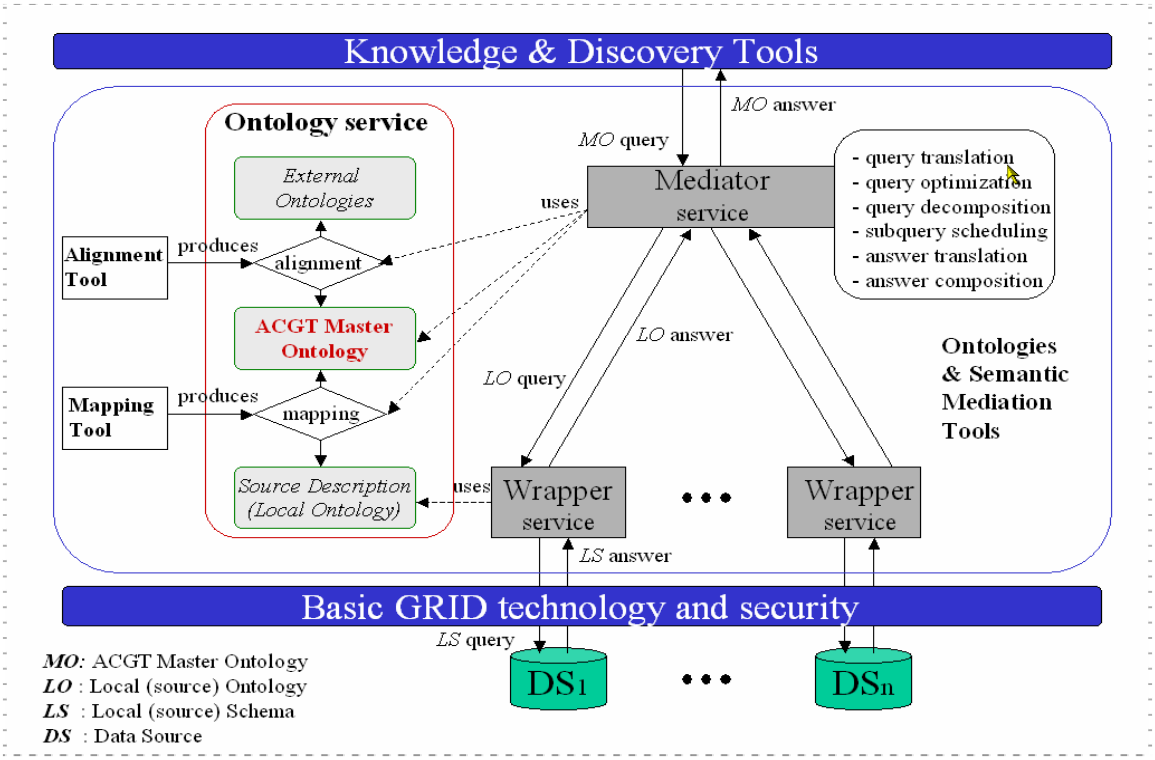


Figure 2: Heterogeneous data integration in *ACGT*

Our approach to heterogeneous data integration is based on a *mediator-wrapper* architecture enabled by the use of ontologies/metadata for the envisioned *ACGT* integration architecture.

In particular, the *mediator* will integrate heterogeneous data sources (which can be *ACGT* databases, external databases, web sources, web data services) by providing a virtual view of all this data. Users (including *ACGT* tools or services) forming queries to the mediated system do not have to know about data source location, schemas, or access methods, since the system will present one shared mediator ontology (the *ACGT* ***master ontology***) to the user and users will form their queries using its terms.

In order for the mediator to integrate the various heterogeneous data sources, their object models, terminologies, embedded domain ontologies, hidden semantic information, query capabilities, and security information will be analysed. Based on this analysis, a *source description* will be defined consisting of a *local ontology* along with a set of metadata, specifying query capabilities and security information. Thus, a source description is an abstraction of a particular data source, possibly conveying semantic information not present in the data source schema.

3.3 An Ontology for Clinical Trials on Cancer

The huge amount of heterogeneous data that genomic and epidemiological researchers share has generated important challenges for information access, integration and analysis that biomedical informaticians must address.

In biomedicine, there are no current ontologies that integrate both genomic and clinical data as they are actually needed in topics such as information retrieval or data mining. For these expectations to be achieved, more domain ontologies should be developed in biomedicine. While there are great expectations for the future, with extended opinions proposing to develop large biomedical ontologies, there is also a need for pragmatism.

Currently, the ontologies (or vocabularies, more strictly) used (such as the UMLS, which now includes Gene Ontology) in biomedical informatics are plagued with technical problems that need to be solved by software engineers. However, there is a vast amount of real problems that can benefit from using these ontologies or vocabularies. For instance, designing new models for biobanks, i.e. databases that include both clinical and genomic data, the unification of databases including information such as Single Nucleotide Polymorphisms (SNPs), using clinical data in drug discovery, improving data mining and searching, and many others.

It is doubtful that ontological research will have a significant impact “per se” in achieving outstanding scientific advances in biomedical informatics. To have realistic chances of success, it will need to link achievements in ontological research to BMI methods and procedures, as well as to consider and address actual BMI research issues [13].

The key semantic integration architectural objectives in *ACGT* include:

- the development of semantic middleware technology, enabling large-scale (semantic, structural, and syntactic) interoperation among biomedical resources and services on an as-needed basis;
- the development of a shared mediator ontology, the *ACGT* Master Ontology, through semantic modeling of biomedical concepts using existing ontologies and ontologies developed for the needs of the project;
- the mapping of local conceptual models (clinical, genomic) to the shared ontology while checking consistency and integrity of the mapped information;
- the development of a semantic-based data service registry to allow advertisement and discovery of data services on the grid. Such a registry will allow *ACGT* clients to discover data services that have a particular capability or manage a particular data source;
- the semantic annotation and advertisement of biomedical resources, to allow metadata-based discovery and query of biomedical resources by users, tools, and services;
- the descriptions of wet lab experiments, in silico experiments, and clinical trials augmented with metadata so as to provide adequate provenance information for future re-use, comparison, and integration of results.

3.4 eScience Workflows.

In providing an open, integrated environment for Clinical Trial management using workflows, ACGT will need to accommodate/integrate a vast range of resources in terms of data and applications. These resources may be within an organisation, for example in-house systems at a given clinical research organisation or local tools developed within an academic research group, or may be external services delivered by a public body or accessed across an extranet.

The European Bioinformatics Institute (http://www.ebi.ac.uk/services/index.html) alone hosts over 50 tools and 40 databases.

The *ACGT* project has identified key user needs wrt to clinical trial workflows. These are:

- *Workflow lifecycle:* Use of a workflow as part of a scientific endeavour requires support for the workflow lifecycle.
- *Semantic description of workflows:* The workflows (and resources) for a particular clinical trial will not necessarily be known a-priori. Specification at a semantic level of the resources and activities required will allow dynamic discovery of suitable resources (in the context of a European open federation of resource providers and resource consumers) and workflows.
- *Workflow provenance:* Use of workflows as part of scientific activity often require provenance data [14] to be kept about activities performed during workflow execution (e.g. details of specific service providers, versions of data and tools involved, etc).

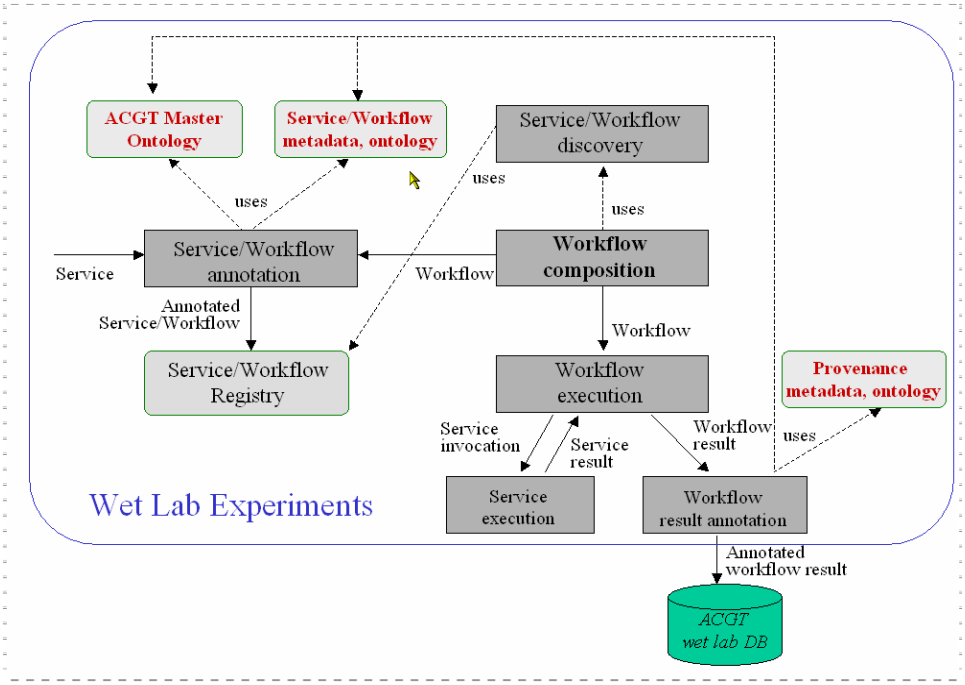


Figure 3: The use of ontologies and metadata in wet lab experiments

The *ACGT* master ontology, along with additional service/workflow metadata and ontologies, will also be used for annotating services and ready made workflows (involved in wet lab experiments). Service and workflow annotations will provide information regarding the service interface, functionality, provider, quality of service, etc. Annotated services and workflows are registered in the service/workflow registry, organized in classes. The use of ontologies and metadata in wet lab experiments is graphically shown in Fig. 3. Based on these annotations, and assisted by the service and workflow discovery module, the user should be able to semi-automatically compose new scientific workflows.

4. Conclusions

ACGT brings together internationally recognised leaders in their respective fields, with the aim to deliver to the cancer research community an integrated Clinico-Genomic ICT environment enabled by a powerful GRID infrastructure.

In achieving this objective *ACGT* has formulated a coherent, integrated workplan for the design, development, integration and validation of all technologically challenging areas of work. Namely: (a) **GRID**: delivery of a European Biomedical GRID infrastructure offering seamless mediation services for sharing data and data-processing methods and tools, and advanced security; (b) **Integration**: semantic, ontology based integration of clinical and genomic/proteomic data - taking into account standard clinical and genomic ontologies and metadata; (c) **Knowledge Discovery**: Delivery of data-mining GRID services in order to support and improve complex knowledge discovery processes.

The technological platform will be validated in a concrete setting of advanced *clinical trials on Cancer*. Pilot trials have been selected based on the presence of clear research objectives, raising the need to integrate data at all levels of the human being.

ACGT promotes the principle of open source and open access, thus enabling the gradual creation of a European Biomedical Grid on Cancer.

ACKNOWLEDGMENTS

The authors wish to express their gratitude to the whole of the *ACGT* consortium for their contributions with various ideas on which the *ACGT* project was developed. The *ACGT* project is funded by the European Commission (Contract No. FP6/2004/IST-026996).

REFERENCES

1. Editorial. Making data dreams come true. (2004), Nature 428, 239.
2. I. Foster. The Grid: A New Infrastructure for 21st Century Science. (2002), *Physics Today*, 55(2):42-47.
3. C. Sander. Genomic Medicine and the Future of Health Care. (2000), Science, 287(5460): 1977-1978.
4. D. Fenstermacher, C. Street1, T. McSherry, V.l Nayak, C. Overby, M. Feldman. The Cancer Biomedical Informatics Grid (caBIG™). (2005). Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, Shanghai, China.
5. http://www.nbirn.net/
6. M. Bertero, P. Bonetto, L. Carracciuolo, L. D'Amore, A. Formiconi, M. R. Guarracino, G. Laccetti, A. Murli, G. Oliva. "MedIGrid: A Medical Imaging Application for Computational Grids," (2003), International Parallel and Distributed Processing Symposium; p. 252.
7. http://www.mygrid.org.uk
8. http://www.bcgsc.ca/discoveryspace/
9. G. Kickinger, P. Brezany, A Min Tjoa, J. Hofer, “Grid Knowledge Discovery Processes and an Architecture for Their Composition”. In: IASTED Conference 2004, Innsbruck, Austria, February 17-19, 2004
10. The Globus Alliance. www.globus.org
11. The OGSA –DAI project. www.ogsadai.org.uk
12. W. Sujanski, Heterogeneous Database Integration in Biomedicine. (2001), Journal of Biomedical Informatics;34(4):285-298.
13. J. Köhler, S. Philippi, M. Lange. SEMEDA: ontology based semantic integration of biological databases. (2003), Bioinformatics 19(18):2420-2427.
14. M. Greenwood, C. Goble, R. Stevens, J. Zhao, M. Addis, D. Marvin, L. Moreau, T. Oinn. Provenance of e-Science Experiments - experience from Bioinformatics, (2003), Proceedings UK OST e-Science 2nd All Hands Meeting 2003, Nottingham.