



Specification of scenarios for a range of integrated demonstrators of the ACGT platform

Project Number:FP6-2005-IST-026996

Deliverable id: D13.5

Deliverable name: Specification of scenarios for a range of integrated demonstrators of the ACGT platform

Date: September 2010

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:	D13.5
Document name:	Demonstrator Specifications
Document type (PU, INT, RE)	PU
Version:	0.2
Date:	07.09.10
Authors: Organisation: Address:	ACGT Technical Management Committee Edited by David Bernasconi

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

ABSTRACT: The present document describes the scenarios used for the demonstration of the progresses achieved in ACGT during the final review of the project.

Key words: ACGT demonstrators, clinical scenarios, technological scenarios

MODIFICATION CONTROL			
Version	Date	Status	Editor
0.1	20.08.10	Draft	DB
0.2	03.09.10	Draft	DB

List of Contributors

- Alberto Anguita, UPM
- David Bernasconi, SIB
- Erwin Bonsma, Philips
- Francesca Buffa, U Oxford
- Stefan Castille, Custodix
- Evangelia Daskalaki, FORTH
- Christine Desmedt, IJB
- Martin Doerr, FORTH
- Maximiliano Garcia, UMA
- Norbert Graf, UdS
- Johan Karlsson, UMA
- Aran Lunzer, UHok
- Luis Martin, UPM
- Juliusz Pukacki, PSNC
- Javier Rios, UMA
- Stefan Rüping, FhG
- Fatima Schera, FhG
- Thierry Sengstag, SIB
- Stelios Sfakianakis, FORTH
- Georgios Stamatakos, ICCS/NTUA
- Oswaldo Trelles, UMA
- Manolis Tsiknakis, FORTH
- Dennis Wegener, FhG
- Gabriele Weiler, FhG

1	SIOP SCENARIO.....	7
1.1	Overview.....	7
1.2	Data pools used in the scenario.....	8
1.3	Providing access to resources.....	8
1.4	Dynamic deployment of services.....	9
1.5	Mediation process.....	12
1.6	Analysis pipeline.....	12
1.7	Potential extensions of the scenario.....	12
2	MCMP SCENARIO.....	14
2.1	Introduction.....	14
2.2	Connection to the ACGT portal.....	17
2.3	Workflow and Service search (recommendation).....	17
2.4	Workflow customization.....	17
2.5	Workflow execution (Dynamic scheduling of R).....	17
2.6	Literature Mining.....	17
2.7	Result browsing.....	17
3	TOP SCENARIO.....	19
3.1	Introduction.....	19
3.2	Data pools used in the scenario.....	20
3.3	Analysis pipeline.....	20
3.4	Analysis output:.....	21
4	APPENDIX A: ACGT ARCHITECTURE VERSION 2 REV 8.....	22
5	APPENDIX II: LIST OF ACRONYM.....	23

Executive Summary

The present document describes the various scenarios used for the demonstration taking place during the final review of the project ACGT in September 2010

The scenarios retained for the demonstration have been selected to illustrate novelties in the development of the ACGT environment. Following the general data flow from clinical trial data collection to data mining tools the following scenarios are described in the present document:

- MCMP scenario: This scenario was retained to demonstrate new features of the data mining environment, such as integrated literature mining tools, GridR scheduling, service discovery, etc.
- SIOP scenario: An integrated scenario bringing data queried from an ObTiMA connected database to the ACGT data mining environment, through the data access and mediation layers. The use of genuine ObTiMA database illustrates the semantic mediation process with full complexity. Dynamic deployment of data access services (DAS) will be demonstrated in this context.
- TOP scenario: demonstrates the ability of the tool to work with various types of genetic data (gene expression microarray, single nucleotide polymorphism microarrays) in the analysis workflow. It also illustrates the possibility of integration of external services such as command line tools or literature mining tools within a the ACGT analysis environment.

The resources used in the demonstrators are summarized in this section.

ACGT portal: <http://epimetheus.ics.forth.gr:8080/acgt/portal?cid>

Role	Machine	Location
ObTiMA web application	obtima.ibmt.fhg.de	FhG (DE)
ObTiMA CRF repository	obtima.ibmt.fhg.de	FhG (DE)
Clinical database (pseudoTOP)	iapetus.ics.forth.gr	FORTH (GR)
Microarray database (pseudoTOP)	base2.thep.lu.se	Lund U (SE)
Data access services	gridnode.ehv.campus.philips.com	Philips (NL)
Gridge Data Management System (GDMS)	moss1.man.poznan.pl	PSNC (PL)
Gridge Resource Management System (GRMS)	moss2.man.poznan.pl	PSNC (PL)
GridR service	kd-9.iais.fraunhofer.de	FhG (DE)
R execution nodes	kd-9.iais.fraunhofer.de moss3.man.poznan.pl	FhG (DE) PSNC (PL)
Meta-data repository	mango.ac.uma.es	UMA (ES)
Mediator service	servet.dia.fi.upm.es	UPM (ES)
Workflow editor/enactor	iapetus.ics.forth.gr	FORTH (GR)
Gridge Authorization Service (GAS)	gas.custodix.com	Custodix (BE)
Submission Tool services	saruman.ics.forth.gr	FORTH(GR)
MyProxy service	myproxy.custodix.com	Custodix (BE)
User registration	acgt_registration.custodix.com	Custodix (BE)

Table 1: List of computational resources used in the demonstration.

1 SIOP scenario

Coordinator: Thierry Sengstag, David Bernasconi

Partners involved: UdS, UPM, UMA, Philips, SIB, FORTH, Siveco, PSNC, Custodix, FhG, LundU (plus indirect contributions of UOxf and INRIA)

1.1 Overview

In this scenario a research analysis is conducted using data retrieved from the database underlying ObTiMA and from the BASE microarray repository. Clinical data are retrieved from this database using the data access services, via the mediator, thus illustrating for the first time a data flow from the CTMS to the data mining infrastructure. Figure 1: General overview of the SIOP scenario shows the general data flow in this scenario.

Anonymization occurred in a preparatory phase of the scenario and is not demonstrated. In order to stick to the ACGT data access policy, all execution steps will be conducted on behalf of a single user legally entitled to access the data.

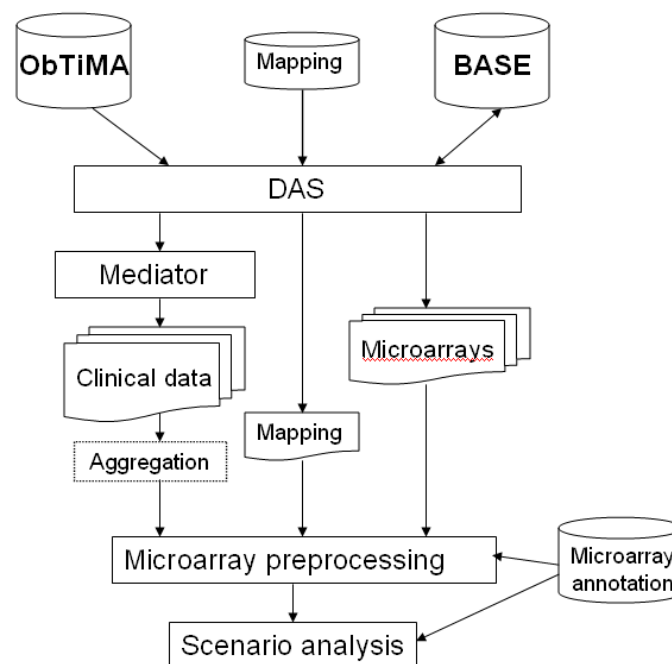


Figure 1: General overview of the SIOP scenario

1.2 Data pools used in the scenario

The data analysis algorithm carried out in this scenario uses data from various sources:

- **Clinical trial data collected using CRFs:** Clinical trial data are stored in ObTiMA: In order to bring together the clinical information needed for the analysis, five mediator queries on the ObTima database are included in the analysis workflow. A R script is used to gather the five queries outputs in order to construct two types of information needed for the analysis:
 - **Generic patient information**, age of patient at time of diagnosis (in days)¹
 - **Survival information**. Relapses data if any, follow-up time
- **Microarray gene expression.** The microarray analysis results of tumour samples taken after surgery. These are stored using the DMS.
- **Mapping table.** A small table that maps the information in the above two databases. As microarray analysis is not a standard part of the SIOP trial, this information is not stored in the SIOP clinical trial system.
- **Microarray annotation** Two files stored in the ACGT DMS, which associate the microarray features (spots) with the corresponding annotation (e.g. sequence ID and gene symbol)

1.3 Providing access to resources

The user *Jane Doe* is owner of the 'public' resources that will be used in this demonstrator. Before the data can be used for the demo access need to be given to the demonstrator user.

Jane Doe carries out the following steps:

1. Log in using the Jane Doe credentials.
2. Open up the GRIDGE Virtual Organisation (VO) Object Managements portlet. Access to a resource can be provided in two different ways, either by assigning an end user to a VO that already has access or by assigning access directly. In this scene access will be granted directly.



Figure 2: VO management tab

¹ The age of the patient at time of diagnosis is obtained by retrieving and subtracting two date values: The date of diagnosis and the patient's date of birth. Due to the anonymisation carried out, the retrieved dates are different from the original (non-anonymised) data. However, for a given patient, the relative differences can still be used for analysis.

3. *Jane Doe* selects the required resource in the resource list, optionally filtering the list to reduce the options.
4. Access to the resource is granted (or denied) by selecting the end user, VO-role or group from the drop-down list. Access granted indirectly (eg. through membership of a group or assignment of a role that has access) can be overridden by specifically denying access. In Figure 3 the end user *Stefan Castille* is denied access, even though the resource is public for all ACGT users.

Jane Doe selects the demonstration user and grants him access to the perform command of the public Wilms resources.

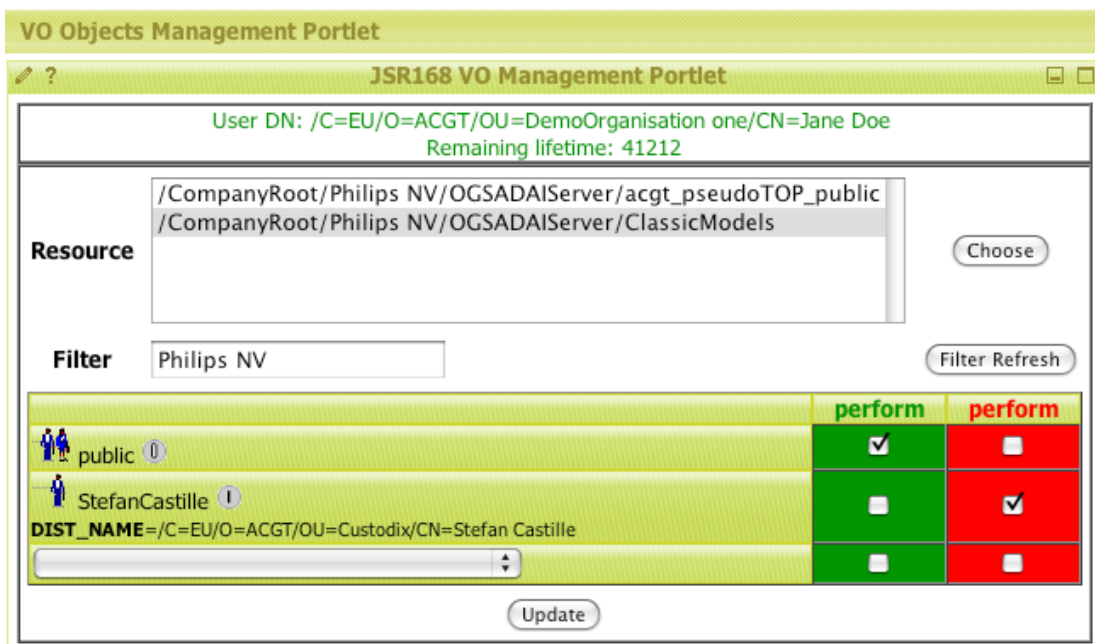


Figure 3: GRIDGE Object VO management portlet

5. Once access is provided for all required resources *Jane Doe* logs out and the demonstration continues.

1.4 Dynamic deployment of services

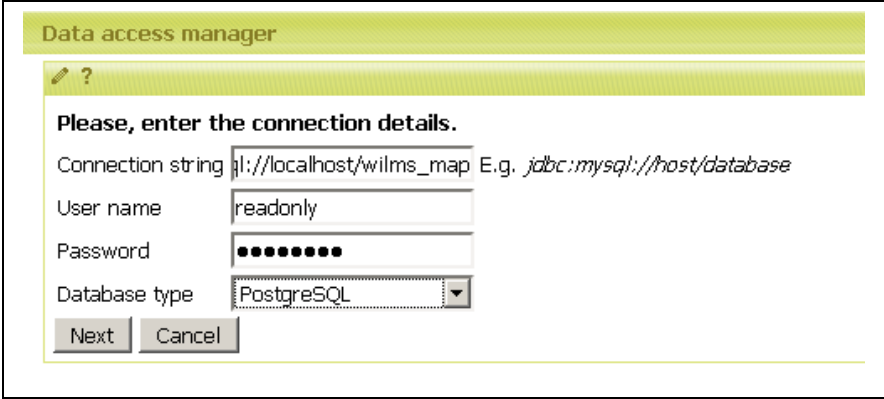
For some data analysis scenarios it can happen that not all needed data is available by way of a data access service. There are various ways this may be dealt with. The data could, for example, be made available as a file, by uploading it to DMS. This is done in the current scenario for the microarray annotation files. When the data is already stored in a database, it may be more convenient to integrate the database that stores the data, so that the data can be subsequently easily extracted using various (potentially complex) queries. Full integration requires amongst others: configuration of the data access web services container, configuration of the GAS, creating mappings at the semantic mediator, and possible minor extensions to the ACGT Master Ontology. This involves changes to various services hosted by different partners, which can take a while. For certain data sources this process is too heavy-weight. Therefore we added a more lightweight mechanism for deploying new data sources, which we demonstrate below by integrating the relational database that stores the mapping table.

The following steps are carried out:

1. Create the mapping to RDF. In order to query the database using SPARQL, the relational schema has to be mapped to RDF. This can be done using the "Data

Access” portlet. The user needs to supply the information needed to access the database (the address of the database, and credentials for access) and an automatic mapping will be generated. (See Figure 4)

2. The user can optionally improve the mapping (e.g. to hide the use of foreign keys)
3. The user deploys the dynamic resource by providing the mapping and a prefix for the name of the resource (Figure 5). A dynamic data access service is deployed for the data resource, and the name of the resource is returned to the user (Figure 6:). The user can use this name to query the resource, and also to remove it when it is not needed anymore².



Data access manager

Please, enter the connection details.

Connection string E.g. *jdbc:mysql://host/database*

User name

Password

Database type

Figure 4: Creating a default mapping to RDF for the Mapping database.

² Only the user that created the resource is allowed to remove it and credential-based authentication is used for this.

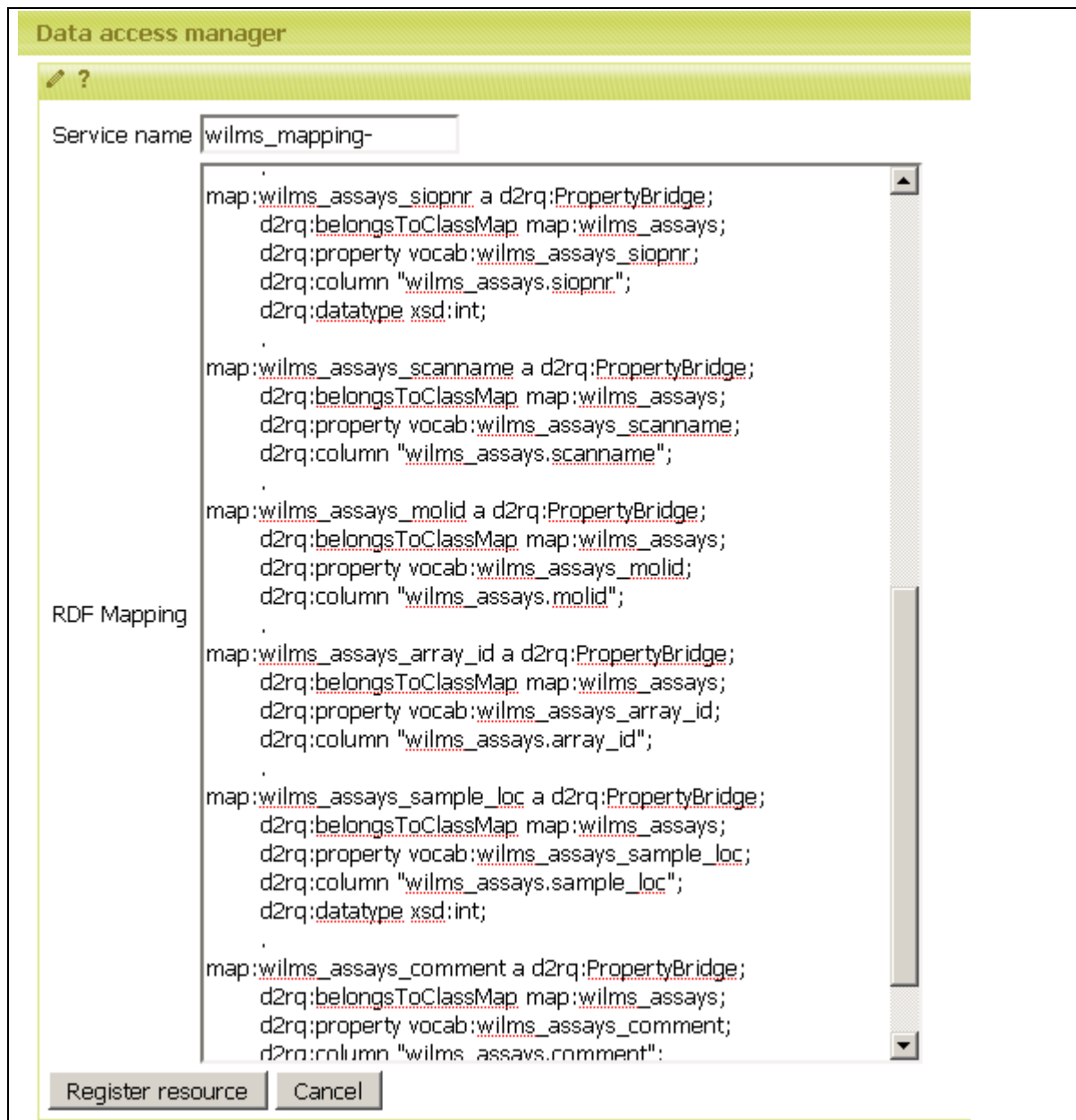


Figure 5: Creating a dynamic data service for the Mapping database, using the provided mapping to RDF.

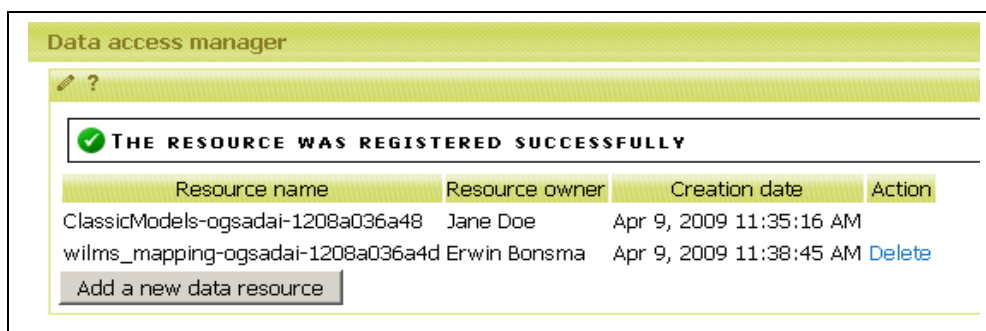


Figure 6: Successfully created the data service. It is now ready to be used.

1.5 Mediation process

This scenario involves the majority of the semantic mediation tools. A set of mappings for the SIOP database has been established using the mapping Java API, thus allowing the query of SIOP database. Once these mappings are completed, they are uploaded to the mediator using the new mapping updating service. The semantic mediator establishes communication with the SIOP data access service to retrieve information after receiving a query that is triggered by ObTiMA.

In the present scenario, demographics information is queried using the full mediation infrastructure of ACGT. However attempting the retrieval of follow-up-related patient information (such as relapse and survival time) yields complex queries, which have performance issues. The latter are under investigation at the time of writing.

The query tool is employed by Obtima users—who access it using the portal—to build the set of queries used in the trial. The tool offers an easy way to build a basic query that can be modified and imported into Obtima for the subsequent access to the mediator.

1.6 Analysis pipeline

In the scientific analysis pipeline, we attempt to reproduce the findings obtained in a research paper obtained using Wilms-tumor samples collected in the SIOP trial and hybridized on two-colors microarrays.³ The specificity of the scenario is that the clinical data are retrieved directly from the clinical database, allowing to easily repeating the analysis using the latest available results in the follow-up.

The outcome of the initial analysis was lists of genes ordered by their significance in the association with clinical parameters such as relapse, histological risk, patient survival and metastases. In this scenario we tried to find a list of gene whose expression was associated with relapse. As in the information contents of the database evolved in time (e.g. through longer-time patient follow-up and data curation), it cannot be expected to observe rigorously identical results. The correlation between the ordered lists of genes as obtained in the analyses conducted three years apart can be assessed through gene-set enrichment analysis (GSEA) and the associated statistical test. Thus the primary scientific outcome of the present scenario is an independent validation of the results found in the original article.

1.7 Potential extensions of the scenario

Two natural extensions of the scenario would be:

1. Use of the external resources provided through BioMoby to create a dynamic annotation of the microarray using the latest official nomenclature for gene symbols, and to take into account the latest clone annotation. (About a third of the clones had their annotation changed since 2006, the date of the annotation of the chip in the article.) Execution of such external services will be demonstrated in the Pathway scenario.
2. Identification of relevant literature for the genes identified in the analysis; such literature mining will be demonstrated in the MCMP scenario.

³ B. Zirn et al., Expression profiling of Wilms tumors reveals new candidate genes for different clinical parameters, *Int. J. Cancer*, **118**, 1954-1962 (2006)

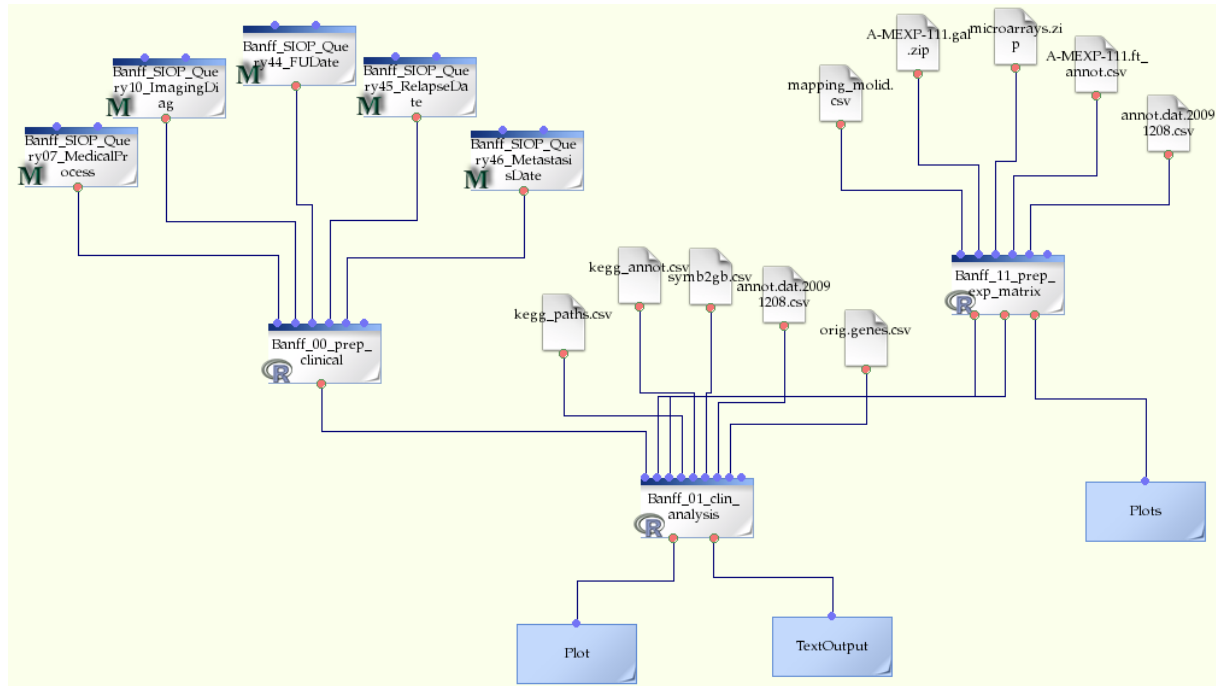


Figure 7: SIOP analysis SIOP workflow

2 MCMP scenario

Coordinator: Dennis Wegener + Francesca Buffa

Partners involved:

FhG, UOxf, IJB, Custodix, FORTH, Biovista, UMA, PSNC, Philips, SIB

2.1 Introduction

Since the last review of the project, numerous improvements have been made to the usability of the ACGT system. In addition, several technical improvements have been made to a number of services and system components. In this part of the demonstration, we demonstrate these improvements based on the known, from the December 2008 review, "multi-center multi-platform", or MCMP scenario. In this scenario a workflow combines Illumina results with Affymetrix results based on real data (PartB of the previous demonstrator, see Figure 8: MCMP workflow). This workflow is extended in the present context to demonstrate novel technological developments that occurred recently in the project and which are described below.

Summarizing the research background of the scenario, it is assumed that biopsies are collected from patients registered in two centers and that each center is using a different microarray platform, namely Affymetrix and Illumina, to measure genes expression in the samples. In addition, the classical clinical parameters associated to each patient are available in a relational database. This specifies a hypothetical multi-centric and multi-platform study, with only one microarray per patient.

In reality, in the present scenario the RNA samples issued from 73 patients have been hybridized on both Affymetrix and Illumina platforms. This unique dataset will allow testing the feasibility of multi-centric multi-platform studies using the high-performance environment of the ACGT connected grid. The full scientific description of this exercise is described in deliverable D12.6.

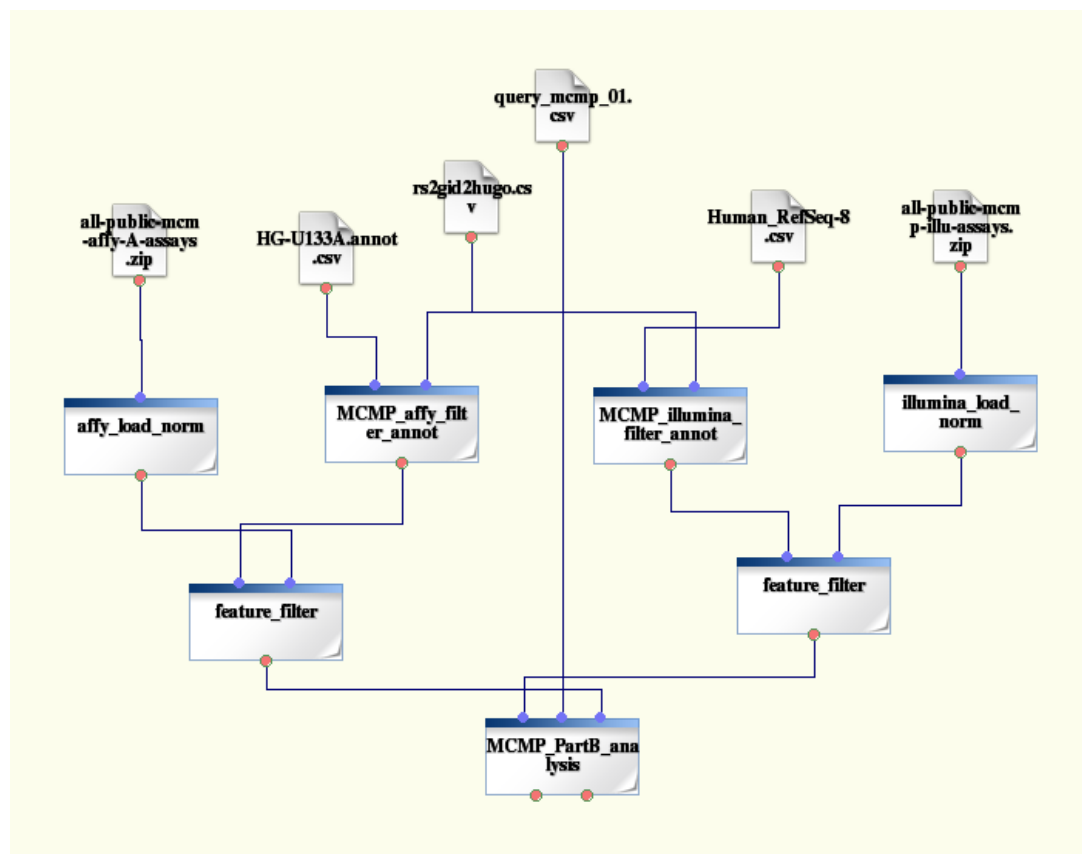


Figure 8: MCMP workflow

Login procedure. The procedure of delegating a credential and logging on into the portal is improved and now consists of only one single step. The improved login procedure is as follows:

The user connects to the ACGT Portal: <http://rd.siveco.ro/acgt/portal> (or the development version <http://rd.siveco.ro:8080/acgt/portal>), selects the "Register now!" tab and uses the Delegation Applet to delegate his/her credentials to MyProxy.

If the delegation is successful the applet automatically logs in the user into the Portal. In addition, if the user logs in for the first time and the credentials are valid, the user account will be created automatically.

Workflow and Service discovery. Discovery of services and workflows can be a complex task when the set of services and data types is sufficiently large. There are real examples of this situation; the BioMOBY Central service catalogue contains over 1600 services and 800 data types.

The service search functionality of the ACGT portal has been improved by using functionality from a new software library (Magallanes). The user interface is similar but important improvements have been made to simplify service and workflow discovery. Users introduce a search string which is compared with various service metadata; name, descriptions, information in external links etc. Furthermore, comparisons are not exact but use a text distance measurement called the Levenshtein distance. This enables the service search component to include services where service descriptions contain similar words to the user query. If no matches are found, the search portlet will display suggestions in the form of "Did you mean...?". The search results are also ranked (i.e. sorted) according to "feed back"-

values which represent the historical record of user selections (the "popularity" of the keywords/resource combination). The feedback value for a search string/resource tuple is increased when a user selects a specific resource and decreased when a user selects another resource using the same search string. It is possible to adjust the decay value that adjusts the feedback values. For example, in a shared environment with several users, the changes in the value should be smaller than in those cases where there is only one user.

This functionality will be demonstrated as part of the MCMP scenario where a service search will be performed. A resource with low ranking will be selected and the following higher ranking of the service will be demonstrated (i.e. the resource is displayed closer to the beginning of the list of results).

Scheduling. So far, the execution of R tasks on the ACGT grid infrastructure was based on providing the information on which machine is capable of executing the R code manually. Now the scheduling components of the grid infrastructure have been improved by allowing dynamic scheduling of jobs according to resource definitions provided by each individual machine. On the one hand, each R script is analysed for collecting the information on which R libraries are necessary for the execution. On the other hand, each machine of the testbed provides the information on which R libraries are installed. Based on this information, the grid middleware finds an appropriate matching of resources.

Literature Mining. We extend the MCMP demonstration by a new workflow component for literature mining. Biovista's literature mining platform involves a number of technical novelties that make it a very powerful resource in the hands of researchers. The main ones are:

- advanced information extraction and gene name disambiguation based on context and SVM (Support Vector Machines)
- support for queries that are oriented towards discovery (such as combinatorial reference analysis) and 'bridge the gap' functions
- no support of NLP directed extraction which means that all relationships that are potentially relevant are considered
- over 25 classes of biologically relevant concepts (genes, diseases, pathways, anatomical locations, PTMs, drugs, etc) that can be searched and semantically correlated

Meta Data. The concept of file related meta data was introduced into the ACGT environment. The main components of the environment that work with files have been adapted to support the specification of a content-type (in the standard MIME format) which describes the format of a file:

- The meta data schema for the file related meta data in the DMS allows storing the content type of a file
- The meta data repository allows for storing a content type for services which work on files
- The DMS portlet allows for adding a content-type to a file uploaded by the user (manually or automatically)
- The GridR service uses the content type meta data for checking if a file provided as input to an R script is of the right format
- The GridR service automatically attaches the content type for output files

In the following, we provide a scene-by-scene description of the scenario describing the various technological components involved and emphasizing which are the novelties with regards to the previous demonstrators.

2.2 Connection to the ACGT portal

The user log on at the portal. The login procedure (delegating a credential and logging in) now consists of only one step.

2.3 Workflow and Service search (recommendation)

The user exploits the search functionality for discovering an existing workflow (the one developed for the Dec. 2008 review). In addition, the user searches for a new service (Literature Mining) which he want to use in order to extend the workflow.

2.4 Workflow customization

The user modifies the workflow by adding the services for the Literature Mining.

2.5 Workflow execution (Dynamic scheduling of R)

The workflow is executed and monitored. At the monitoring portlet, the user can check the job description and see the resource specification for the libraries of the R script which was used for dynamic scheduling.

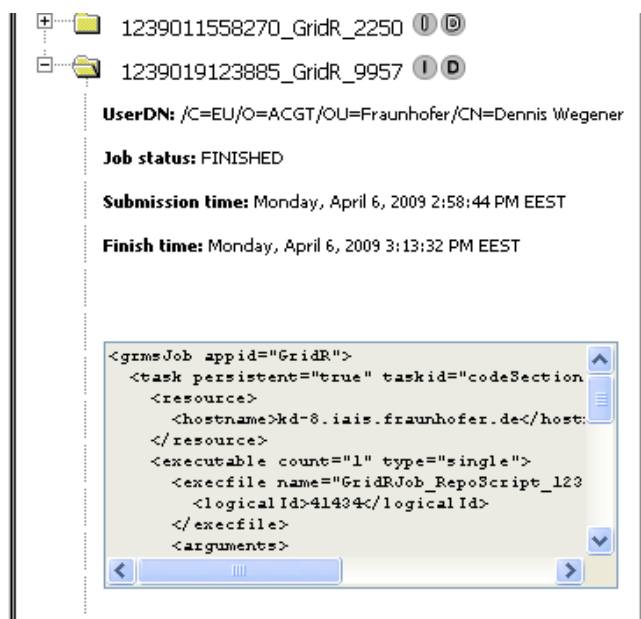


Figure 9: Screenshot - GRMS job description including resource specifications.

2.6 Literature Mining

An advanced service for literature mining is executed as new part of the workflow.

2.7 Result browsing

The final output of the MCMP workflow will include an additional output file containing a list of genes which is used as input for the Literature mining service. The results of the workflow execution can be browsed at the DMS portlet. The user can check the content type of the output files from the GridR executions.

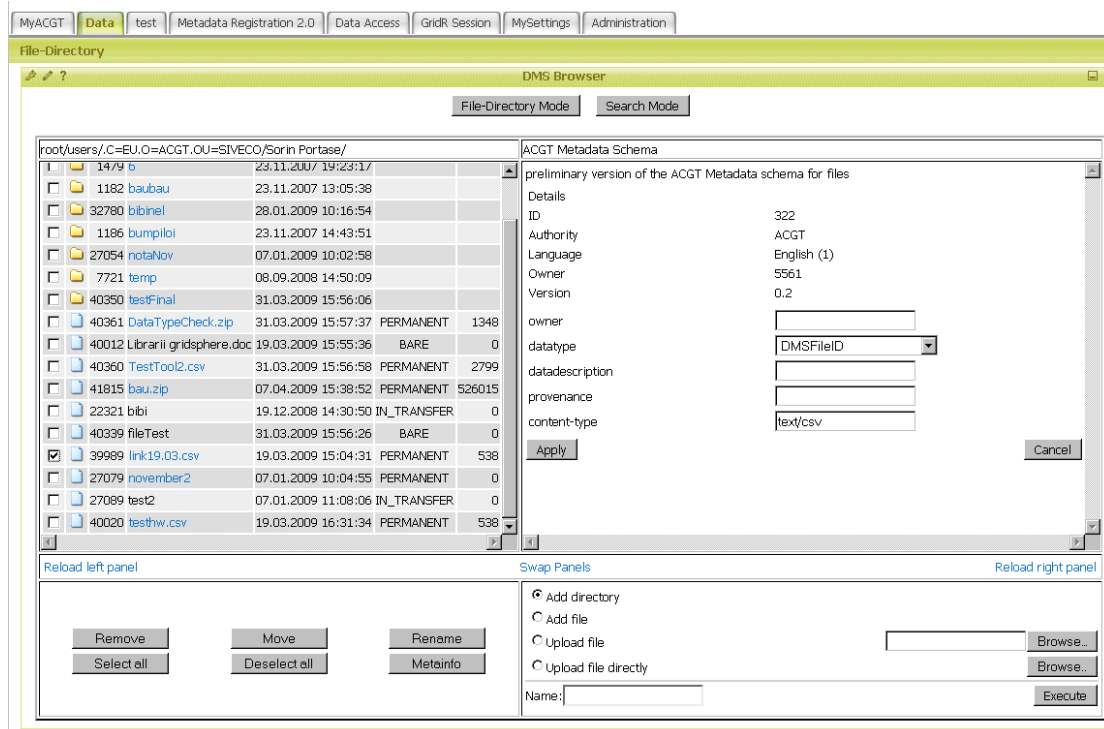


Figure 10: Screenshot - Editing and checking file related meta-data in the DMS portlet

3 TOP scenario

3.1 Introduction

In this scenario, a research analysis is conducted on the data from TOP clinical trial. The TOP trial is designed to prospectively evaluate the predictive value of TOP2A and to identify markers of response/resistance to preoperative epirubicin in estrogen receptor-negative (ER-) breast cancer patients.

The scientific objective of the TOP scenario is to find a list of genes whose expression is significantly associated with response/resistance to the epirubicin pre-operative treatment and verify the predictive capacity of TOP2A as a marker for response/resistance.

In order to achieve this objective, data from 65 patients from the TOP trial are analyzed. Genes identified as significantly associated with response are then compared with relevant publications through the literature mining tool.

In an ACGT's point of view, the objective for this scenario is to demonstrate the ability of the tool to work with various types of data. Indeed, this is the first scenario using SNP microarray data in the analysis workflow. It also illustrates the possibility of integration of external services such as command line tools or literature mining tools within the ACGT analysis environment.

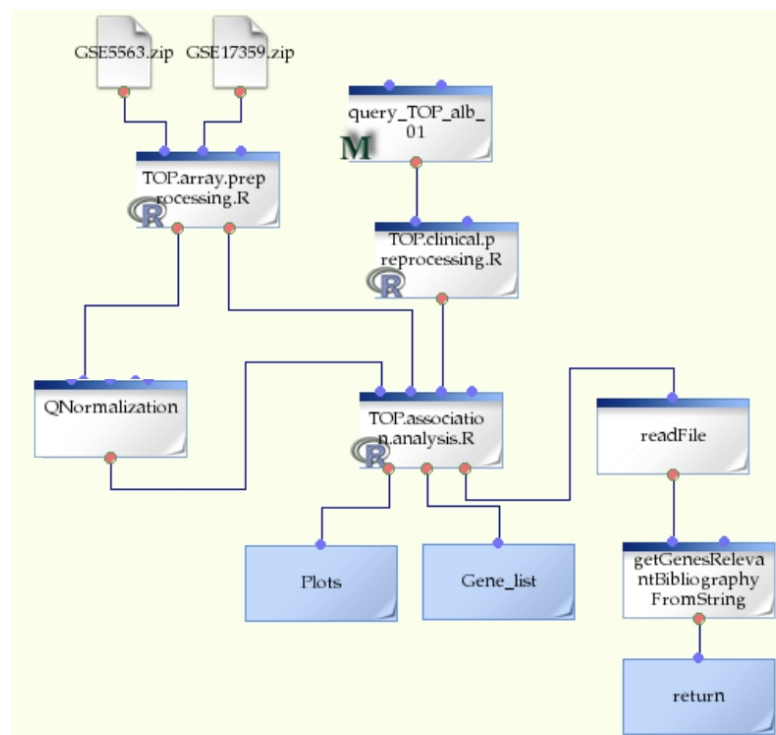


Figure 11: The TOP scenario's workflow

3.2 Data pools used in the scenario

The scenario takes advantages of three types of data, namely clinical data, gene-expression microarray data, and SNP (single nucleotide polymorphism) microarray data.

- **Clinical trial data accessed via mediator queries:** For this scenario, 7 queries are needed for constructing the clinical information matrix. The fields required for the analysis are the following:
 1. Patient's identifier in the trial
 2. Patient's birth data
 3. Patient's diagnosis date
 4. Tumor's histopathologic grade
 5. Tumor's T classification,
 6. Tumor's N classification
 7. End of treatment reason
 8. Pathological complete response status

- **Gene-expression microarray data retrieved from DMS:** Gene-expression information come from an Affymetrix GeneChip Human Genome U133 Plus 2.0 array and are stored in the DMS in compressed zip directories

- **SNP microarray data retrieved from the DMS:** SNP information come from an Affymetrix Genome-Wide Human SNP Array 6.0 and are stored in the DMS in compressed zip directories.

3.3 Analysis pipeline

In the scientific analysis pipeline, we attempt to find a list of genes whose expression is associated with response/resistance to the preoperative epirubicin treatment using clinical data and gene-expression data from 65 patients of the TOP clinical trial. The specificity of this scenario is the inclusion of SNP (single nucleotide data) in the analysis process. These data are used to assess the association between copy number variant and the response of patients to the treatment, The second specificity of the scenario is the integration of external services in the workflow.

In order to fulfill the analysis, the workflow is composed by four R scripts:

- **micr.preprocessing.R:** In this R script gene expression microarray data are prepared for the further analysis steps. Microarray .CEL files are stored in a compressed directory. The R script download, uncompress and store the microarray data in the current working directory. Gene-expression intensities then are computed

but not normalized as this step is performed by an external command line tool. (QNormalization box in Figure 1).

- **snp.preprocessing.R:** In this R script, single nucleotide polymorphism microarrays are prepared for the further analysis steps. The snp .CEL files are stored in the DMS in a compressed directory. The R script download, uncompress and compute for each probeset the copy number as well as the location of the probeset in the genome. The output of the script is a matrix containing these two information in a .csv format.
- **TOP.clinical.preprocessing.R:** This R script prepare the clinical informations gathered by mediator queries under the form of a matrix in order to be used in the following steps of the analysis.
- **TOP.association.analysis.R:** This R script investigates the association between the genomic data and the clinical data. Inputs of this R script are microarrays normalized intensities as well as clinical data. Outputs of the analysis are reported on point 3.4.

The workflow also contains two external services: a command line tool and a literature mining tool. These tools not only illustrate the capacity of integration of third party tools within the ACGT environment but also gives insight of the relevance of our results by comparing our list of genes with the relevant publications.

1. **The command line tool:** A web service executing a QNormalization of the gene-expression microarrays using several processors (MPI library).
2. **The literature mining tool:** A literature service from Biovista returning the publications referencing one or more of the genes given as input. The first 100 results are returned.

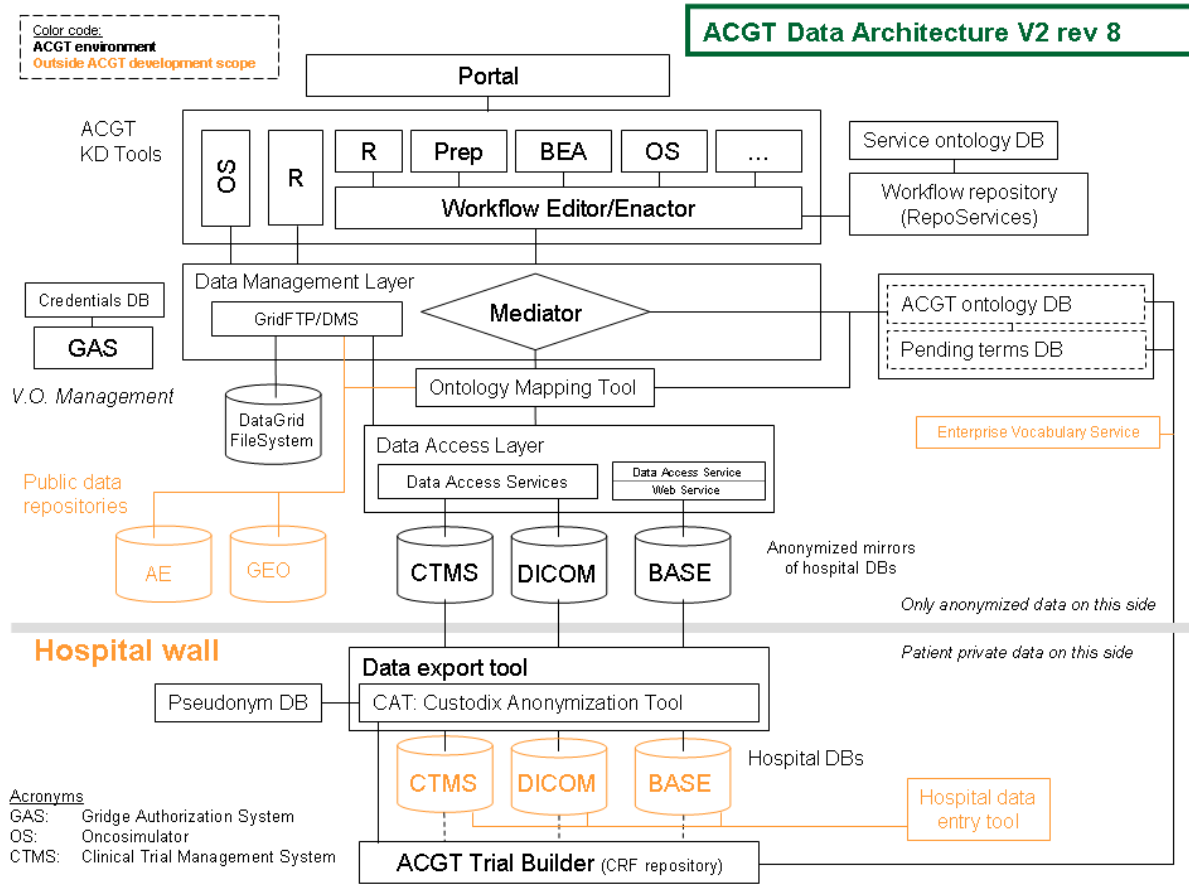
3.4 Analysis output:

Output of this analysis workflow are the following:

- List of genes whose expression is significantly associated with response/resistance to the treatment. The genes are displayed in a html file and ordered according to the significance of their association with the response. For each gene reported, a log₂ ratio of the CNV variant is presented in order to assess the relation between gene-expression and gene copy number.
- Heatmap of the identified genes and the clinical outcome.
- Gene copy number comparison between good and bad responders. For each chromosome, log₂ ratios of SNP copy numbers are reported in plots. The plots allows the identification of chromosomal region where copy number is different between patient with good response to the treatment and patients with bad response to the treatment.
- Literature associated with identified genes. The final workflow's output consists in a list of publication related with the list of genes identified as associated with the response to the treatment. The list of publication contains hyperlink to the Pubmed database.

4 Appendix A: ACGT architecture Version 2 rev 8

The figure below presents the reference ACGT architecture.



5 Appendix II: List of acronym

API	Application Programming Interface
ACGT CA	ACGT Certification Authority
CRF	Case Report Form
DBMS	Data Base Management System
DAS	Data Access Services
DMS	Data Management System
GAS	Gridge Authorization Service
GDMS	Gridge Data Management System
GRMS	Gridge Resource Managment System
MO	Master Ontology
OWL	Web Ontology Language
RDF	Resource Description Framework
SPARQL	SPARQL Protocol and RDF Query Language (recursive acronym)
VO	Virtual Organization