Introduction of a new clinical trial in ACGT: a case study

Project Number:	FP6-2005-IST-026996
Deliverable id:	D7.8
Deliverable name:	Introduction of a new CT in ACGT: A Case study
Submission Date:	December 2009

Information Society Technologies

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 7.8
Document name:	Introduction of a new CT in ACGT: A Case study
Document type (PU, INT, RE)	PU
Version:	1.0
Submission date:	11/06/2010
Editor: Organisation: Email:	Luis Martín UPM lmartin@infomed.dia.fi.upm.es

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

ABSTRACT: This deliverable aims at describing the experience of the inclusion of a new CT (at its early stages) in the ACGT platform. In this particular case, we despict an experimental scenario where data sets already exist before the begining of the trial. More concretely, the experiment deals with the integration of microarray (Arrayexpress) and relational clinical (SIOP) data. This document shows the process of setting up the new trial in the platform, describing the different components involved, and despicting results of the experiments carried out.

KEYWORD LIST: Clinical trials, Semantic Mediation, Wrapper, Microarray data

MODIFICATION CONTROL			
Version	Date	Status	Editor
0.1	10/07/2009	Draft	Luis Martín
0.2	25/11/2009	Draft	Luis Martín
1.0	11/06/2010	Final	Luis Martín

List of Contributors

- Alberto Anguita, UPM
- Luis Martin, UPM
- Thierry Senstag, SIB
- Stefan Rueping, FHG
- Norbert Graf, SAAR
- Stelios Sfakianakis, FORTH
- Gabriele Weiler, IBMT-FHG
- Martín Esteban, UPM
- Diana López, UPM

Contents

1 Executive Summary

This deliverable aims at describing the experience of the inclusion of a new CT (at its early stages) in the ACGT platform. In this particular case, we despict an experimental scenario where data sets already exist before the begining of the trial. More concretely, the experiment deals with the integration of microarray (Arrayexpress) and relational clinical (SIOP) data. This document shows the process of setting up the new trial in the platform, describing the different components involved, and despicting results of the experiments carried out.

2 Introduction

Purpose and structure of this document

This document aims at describing the process of integrating a new Clinical Trial with already existing data sets in the ACGT platform. Section 3 is devoted to the design and implementation requirements. Section 4 despicts the different technical components involved in the process. Section 5 shows the methodology for the inclusion of this trial in the platform, and section 6 concludes explaining the suitability of the ACGT approach in this particular case.

Introduction

Clinical trials conducted by researchers often allow extracting valuable knowledge about treatments for diverse diseases. In these trials, different patient's data is recorded for subsequent analysis. Data mining techniques allow clinicians to come up with useful information about which treatments better suit each patient suffering from a specific disease. These data usually covers the patient's clinical data. Some clinical trials however have proven that in many cases, the patient's clinical data is not enough to obtain a suitable treatment, given that different patients with similar clinical characteristics present dissimilar response to the same treatment. More biological information in these cases permits a more subtle distinction between the trial patients. This extra information often has its provenance in the post-genomic data. Modern analysis techniques allow extracting the gene expression signature of a patient, adding a great value to the classical and already available clinical data. Numerous studies employing this kind of data have allowed establishing the relation of the expression of a gene (or lack of expression) in a patient with his predisposition to suffer a disease or to answer positively to a specific treatment. In [Petrik et al. 2006], Petrik shows the relation between genomic signature and brain tumors. Ippolito proved as well the relation between gene expressions and human neuroendocrine cancers outcomes [Ippolito et al. 2005]. This achievement not only facilitates the discovery of cures for specific diseases, but also enables the possibility of designing patient specific treatments—the so called personalized medicine. The benefits of personalized medicine are manifold. The most obvious, being able to apply the most adequate therapy to each patient based on her biological characteristics. We cannot however ignore the cost savings that this knowledge can imply. Avoiding unnecessary medical tests to patients who are known to not suffer from a specific disease, or select the tests that check the diseases a patient is more keen to suffer should help reduce unnecessary expenses.

The importance of developing personalized treatments is of special relevance in the case of clinical trials on cancer. The harm provoked by an excessive dose of chemotherapy to a patient is far greater that with other treatments. All modern trials on cancer rely on post-genomic data analysis techniques to improve their results. In the end, the expected outcome of a clinical trial is the identification of one or more biomarkers that allows an efficient clustering of patients into several categories. Each category will group together patients with similar responses to treatments, but different from other clusters. Clinicians can then design the most appropriate treatment for each cluster, hence minimizing the patient's risk of receiving non adequate doses. Numerous examples of studies finding correlation of genomic signature and treatment outcomes can be found in literature. Li [Li et al. 2005] describes differences in gene expressions of Wilms Tumors against normal kidneys. In [Zirn et al. 2006], a set of genes associated to relapsed Nephroblastoma tumors in patients treated with chemotherapy was identified. Wang [Wang et al. 2005] identified a 76-gene signature that allowed high precision predictions for metastases development within 5 years in cases of breast cancer.

Still, the use of post-genomic data by itself is not useful. Genomic data must be analyzed in conjunction with clinical data in order to obtain useful results. This leads to the necessity of medical researchers to be able to perform integrated access to these types of data. In some cases, databases have been designed ad-hoc and present no problems for being integrated. Many times, however, these types of data repositories present completely different provenance and structure, making it very difficult for the biostatician to carry out the subsequent data mining techniques. In the end this turns up to be an important bottleneck in the complete data analysis process. It is thus required not only a multicentric approach for data integration, but also a multilevel approach that allows the seamless access of clinical and genomic data together. The ACGT project aims at providing a technological platform that covers this issue by means of a software module that offers uniform access to clinical data and public genomic data.

3 Design and implementation requirements

The ArrayExpress public database

In 2001, a first version of the public database for genomic data called ArrayExpress was released by the MGED group (http://www.mged.org/) [Brazma 2001]. This release included a specification for descriptions of microarray experiments. This description has served as the basis for the development of the MAGE-OM object model, a set of more than one hundred classes describing the domain of microarray experiments. From this model, a new markup language, named MAGE-ML, was created [Spellman 2002]. From this object model and its corresponding markup language, ArrayExpress was created [Brazma 2003][Sarkans 2005]. ArrayExpress can be described as a public compilation of microarray experiments. It allows institutions and laboratories to upload their data to a central repository, and share other's data. It was quickly endorsed by a great number of laboratories, making it the largest microarray experiments community to the data.

Methods

The goal of WP7 in ACGT is to support and offer the technologies that allow seamless integration of data from clinical trials on cancer. For this purpose, we have targeted the inclusion of ArrayExpress data in our platform. The semantic mediation platform in ACGT deals with the semantic heterogeneities that arise in the process of integrating disparate biological sources. The rest of heterogeneities, namely the syntactic ones, were so far tackled by a different layer in the ACGT platform—the so called database wrappers. The database wrappers were developed in ACGT with the goal of hiding the syntactic peculiarities of each specific data repository, in terms of access interface and query language. By means of these wrappers, the semantic mediation layer was able to access every data source with a common web service interface and a unique query language: SPARQL. Wrapper modules were developed in ACGT to access legacy SQL databases and image repositories. ArrayExpress was not however included in the catalog of data sources accessible through a wrapper. Under this circumstance, we decided to develop our own wrapper for ArrayExpress. The choice was mostly motivated by the benefits of maintaining the structure of the semantic mediator intact: in the absence of a wrapper hiding the syntactic characteristics of the ArrayExpress interface the mediator would require deep restructuring. Our wrapper for ArrayExpress would be seen by the mediator like the rest of existing wrappers, thus maintaining compatibility. In conclusion, the problem is reduced to the development of a wrapper for ArrayExpress which features the same characteristics as previous wrappers in ACGT.

The ArrayExpress wrapper

ArrayExpress offers web-based access as well as programmatic access to its repositories. The type of queries that it accepts is however quite simple, and it usually just leads to the retrieval of complete experiments. The structure of the retrieved data on the other hand tends to be rather complicated, differing between experiments, but being always a subset of the MAGE-OM model. In order to fit the ACGT wrapper requirements, our ArrayExpress wrapper should expose an RDF based schema of the underlying data. This schema was manually created using Protégé, by defining a superset that contained all the elements in the analyzed experiments in ArrayExpress. This schema resembles the MAGE-OM model, but is based on RDF, as required in our mediation platform. A separate module was developed in order to automatically download specific experiment files each time a query arrived. It is able to query the ArrayExpress programmatic interface in order to determine the list of experiments that contain data related to a given SPARQL query. Those experiments are downloaded as XML files in MAGE-ML language. Another module is in charge of parsing those files and translating their contents into instances of our own RDF schema. Therefore, for each query an empty RDF repository is created resembling our RDF schema. This repository is then populated using the experiment files as source of information. Once this task is completed, this repository is queried with the SPARQL query received by the wrapper. A web service based interface is added to the wrapper so it offers the same features as the rest of the ACGT wrappers. At this point the only task remaining is to create a mapping from the ACGT Master Ontology—which is employed by the mediator as the global schema—to the RDF based schema created ad-hoc for ArrayExpress. This is a one-time procedure performed by hand—with assistance of our ACGT Mapping Tool. At this moment, the semantic mediator can communicate with the ArrayExpress wrapper, sending it SPARQL queries and receiving results which can be integrated with other repositories integrated in ACGT—i.e. databases from clinical trials. Figure 1 depicts the process for handling SPARQL queries in our ArrayExpress. This process takes place each time the mediator decides a subquery must be sent to the ArrayExpress wrapper.

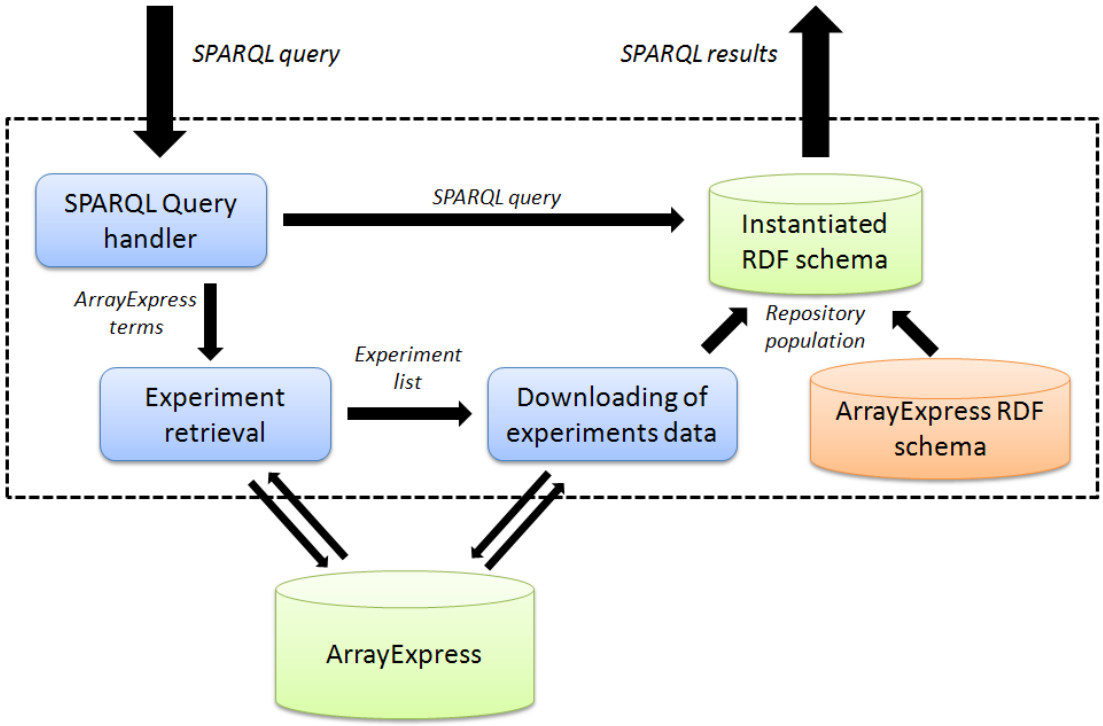

Fig 1. The process of solving SPARQL queries from the mediator involves a set of software modules and the access to ArrayExpress to download the desired data.

4 Description of technical components

This section gives a little description of the different components involved in the ACGT platform, and place them in the context of the present study. Each sub-section includes a table summarizing the dependencies inside the platform and required average effort for a new CT set up.

4.1 Data Access Services

The Data Access Services are the components in charge of homogenizing the access interface to underlying data repositories in order to eliminate syntactic heterogeneities among the disparate data sources that are accessed. This way, client modules of the wrappers (the Semantic Mediator in our case) are offered a unique data access interface. This common interface accepts queries in SPARQL query language, and offers an RDF-based schema for each of the data sources. For each different type of database used in the ACGT platform, a wrapper has been developed. This way, SQL based databases, as well as image repositories are accessible by the mediator in a common manner.

In case of ArrayExpress, there was no wrapper available for the ACGT platform. A new wrapper was developed to provide the required interface to ArrayExpress. However, there were several differences with previous data resources. First, ArrayExpress does not offer a programmatic interface that allows performing any query on its data. It is only possible to perform queries with some keywords as restrictions. Second, previous wrappers translated well-known data schemas into RDF-based schemas (SQL, DICOM…). ArrayExpress uses a proprietary format based on the MAGE-OM object model. For this reason, specific purpose modules were developed in order to fulfill the wrapper specification requirements. Details of this development can be found in section 3.

Component name	Data Access Wrappers
Dependencies	None
Is service for	Semantic Mediator
Requires new software development (MM)	Yes (if it requires access to a new data resource type) (4MM)
Requires data configuration (MM)	No
Requires maintenance	Yes

4.2 Semantic Mediator

The Semantic Mediator is the core component of the ACGT Semantic Mediation layer. It is in charge of accepting queries in terms of the ACGT

Master Ontology and translating them into terms of the physical databases included in the integration platform. The Semantic Mediator can be accessed as an OGSA-DAI service, making it available to any terminal connected to the internet. It offers a SPARQL interface for performing queries in terms of the Master Ontology. The received queries are handled by the Semantic Mediator accordingly to the existing database mappings, so a new query for each un-derlying data source is produced and their results are properly merged and sent back to the user.

For this particular scenario, the Semantic Mediator offers access to a virtual repository representing the integration of microarray and SIOP clinical data. Users are able to launch queries representing integrated data from both sources, or from one of them in isolation. Although the Semantic Mediator requieres no software modifications in the general case, some improvements have been made due to the discoverings made in this case study. The most important one has been the inclusion of dependant cross-reference variables, that requiere of the sequential execution of some queries in order to improve performance (microarray information are normally large data sets of information, so it is recomendable to filter as soon as possible).

Component name	Semantic Mediator
Dependencies	Data Access Services, Master Ontology, Mappings
Is service for	Analytical Tools, Query Tool, Obtima
Requires new software development (MM)	No
Requires data configuration (MM)	Yes, mapping files (2 MM per CT)
Requires maintenance	Yes

4.3 Master Ontology

The role of the ACGT Master Ontology in database integration is twofold, 1) it supports the creation of homogeneous views representing the underlying data sources (the mapping process), and 2) it serves as a vocabulary server to annotate the results of the queries, aiding to generate semantics-compliant result sets. A suitable part of an ontology in a suitable encoding can be used or interpreted as target schema. The MO will be used as our Enterprise or Target Model in order to support the appropriate mappings from our local data schemata (Source models). These mappings will enable the integration under a common knowledge representation model (LAV approach) where data source relations are defined in terms of a global schema.

Mapping specifications should be given by domain experts and should be expressive enough to allow an IT-expert to configure the respective

wrapping and mediation services without further help from the domain expert. A tool is, therefore, required in order to assist the mapping specification process. In order to support a domain expert in the mapping specifications, it is beneficial to mark a layer in the MO which is adequate to the ontological level of detail of characteristic data structures in the domain. Further examples of mappings of characteristic schema constructs can be helpful. It may also be beneficial to mark subsets of the MO by context of application to generate personalized views of the MO.

Component name	Master Ontology
Dependencies	None
Is service for	Obtima, Semantic Mediator
Requires new software development (MM)	No
Requires data configuration (MM)	Yes, ontology updates (2 MM per CT)
Requires maintenance	Yes

4.4 Obtima and the Mapping API

The main goal of ObTiMA is to support the design phase of a clinical trial allowing to set up the Patient Data Management System for a trial in a standardized and userfriendly way integrating the ACGT master ontology into the design process. With the Patient Data Management System (PDMS) an end-user (clinician) is able to manage a patient within a clinical trial, to capture data, to report data and to query the database in a standardized way in terms of the ontology. The seamless integration of the ontology into the design process of a trial guarantees that the data collected during the trial has comprehensive metadata, a crucial condition to leverage the data for further research like cross-trial analysis.

When setting up a new CT, the trial chairman has to set up the new trial (specify trial metadata, CRFs and outline). In principle that should not imply new software development. Regarding the generation of data models or configuration files, there is no need of generating new ones. Only the ACGT Master Ontology has to be extended with the classes and relations necessary for the trial, which are not already included. This can however be initiated by the trial chairman during the process of CRF creation in ObTiMA or via the submission tool.

Component name	Obtima
Dependencies	Semantic Mediator, Data Access Services

Is service for	Analytical Tools
Requires new software development (MM)	No
Requires data configuration (MM)	Yes, mapping files (2 MM per CT)
Requires maintenance	Yes

The Mapping API plays an important role within the ACGT Semantic Mediation layer, as it offers the possibility of including new data sources in the integration platform. This process involves establishing relations between elements of the schema of the source to be inte-grated and elements of the ACGT Master Ontology—which acts as global schema for the mediator. These relations are called mappings, and the process of establishing mappings is called mapping process. The mappings are used by the mediator in the task of translating integrated queries into queries for the underlying databases. The mapping process often requires some degree of expertise on both the domain of the data being mapped and the inner technical characteristics of the mappings being produced. To this end, the Mapping API incorporates a series of features aimed at facilitating this task and reducing as much as possible the user's workload.

Component name	Mapping API
Dependencies	None
Is service for	Semantic Mediator, Query Tool, Obtima
Requires new software development (MM)	No
Requires data configuration (MM)	No
Requires maintenance	No

4.5 GridR

Analytical Tool Services are tools for knowledge discovery and data mining tasks. They perform analyses in input data sets and they are "deterministic" in the sense that when they are given the same data and the same parameters they will produce the same result. The setting up of a new CT would imply the development of new R scripts, though, which is comparable to the development of a new workflow. It is difficult to give an estimate of the effort, because this dependes on the complexity of the scientific question, but, in most cases, it would tale less than 1 PM.

It is assumed that the end users themselves develop the R scripts. They would probably need to communicate with the suppliers of data access

(mediator or other) in order to understand the format and semantics of the data.

Component name	GridR
Dependencies	Semantic Mediator, Data Access Services
Is service for	R scripts
Requires new software development (MM)	No
Requires data configuration (MM)	Yes, R Scripts (1 MM avg)
Requires maintenance	No

4.6 Workflow environment (Workflow editor and enactor)

The ACGT Workflow Environment is a suite of software components where the different services and tools are put together and connected in order to describe and construct an experimental scenario or process. Therefore the inclusion of a new trial has the potential of the need for some additional effort if this new trial requires the integration at the workflow layer some novel analysis tools. Even in this case there's no new development needed if these new tools have been implemented in compliance with the ACGT syntactic and semantic guidelines. If that's not the case we estimate an additional effort of 2 MM on average (because the actual effort needed varies and depends on the way the tools have been implemented).

It is possible that a new service supports a different data model. For example Biomoby services have their own way of serializing their data structures in XML that is not fully type safe and in compliance with the WS-I (Web Services Interoperability Organization) best practices. In this case we estimate an effort of 1 MM (on average again). If the new scenario requires the inclusion of some new tools and/or data management services then we need to interact with the developers of these components In order to become familiar with their tools. In this interaction we need to answer the following questions:

- What is the middleware technology supported by this new tool (e.g. Web Service, HTTP/REST service, command line tool, etc.)?

- What are its security requirements (e.g. does it need Grid (GSI) security credentials)?

- What are the data formats used?

- Does it use Grid data management services for the storage if its output? Does it require the data input to be file references or support passing data "by value"?

- Is it compliant with the metadata repository and its metadata schema for describing its functionality, input, and outputs?

- Are there any other policies and interaction protocols that the users of this tool or the components interacting with it need to follow?

Component name	Workflow editor
Dependencies	All ACGT services
Is service for	Workflow enactor
Requires new software development (MM)	No
Requires data configuration (MM)	Yes (1 MM)
Requires maintenance	Yes

5 Methodology

5.1 Graphical outline of the methodology followed for the integration

Figure 2 shows the general process of integrating new sources in the ACGT platform. The most expensive subprocesses (bottlenecks) are highlighted in red.

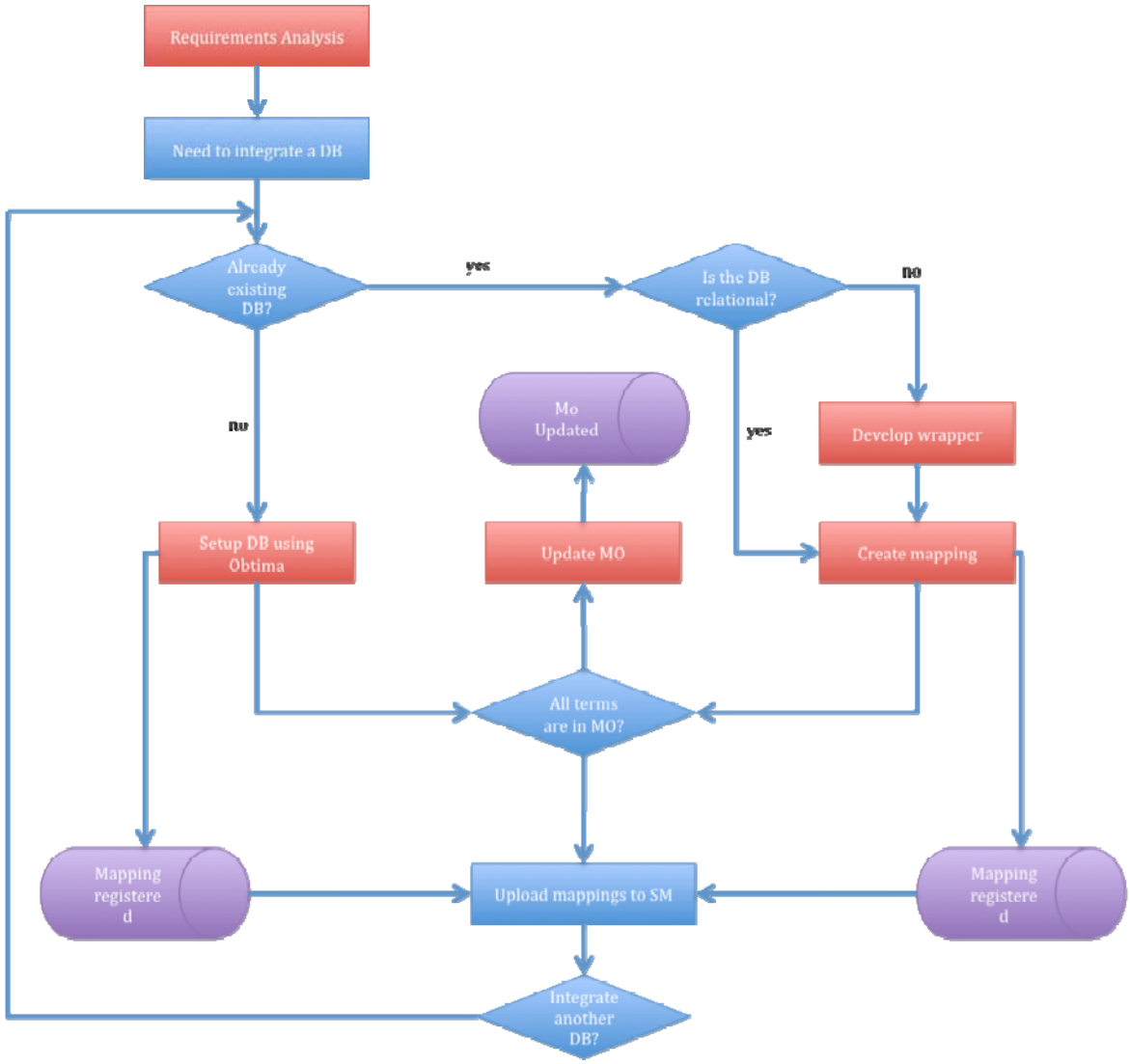

Fig. 2: Methodology for new database configuration in the ACGT platform

6 Conclusions

The ACGT platform is an open technological platform for the end-users to perform their analyses and experiments in the context of state-of-the-art clinical trials. In order to be widely accepted it needs to be sufficiently easy to use, non-intrusive but also proactive to the users' needs and decently efficient in terms of performance and throughput. To this end, the ACGT services and tools and the platform as a whole need to be further polished and tuned.

One of the main components, that is the core interface of the system from the clinicians point of view is Obtima. When setting up a trial with ObTiMA, the main bottleneck is creating ontology-based CRFs. Although this task is probably easier than mapping legacy databases to the ontology and can in principle be achieved by a clinician, it is still tedious work. That is also due to the fact that the user interface for this task in ObTiMA is still in a prototypically state.

The amount of MM is depending on the trial. If it is a completely new cancer one needs more work to put in, as it would be an adult renal tumour. It is depending on the fact if I would be the chairman of the trial or a participant of the trial. It is also depending on the time. At the beginning in building the new trial I do need more time then during the run phase. At the end again more MM is needed, because of clearing and analysis of data. From a clinical side it also important to have the ability to import data from old trials into ACGT or ObTiMA easily. Performance and usability of all parts of ObTiMA will be improved in the future, what are currently probably its main bottlenecks.

From the data analysis point of view, internal experts assert that in general it is easy to use the platform. However, in some cases the help of experts in specific domains is needed, mainly when we talk about define the details of analysis scenarios. This occurs when I have to use "exotic features" (from a standard bioinformatician viewpoint) of the platform, such as creating SPARQL queries on dynamically registered databases, or when I need "backdoor" actions such as when registering a static database to the system.

Develop data analysis workflows around a new scenario, or at least provide advice on how to use the platform usually take an estimated effort of 2 MM.

From a data miner/biostatistician viewpoint, the main bottlenecks will be:

1. The definition of procedures to guarantee the consistency of data identifiers across databases (e.g. if clinical data are stored in a mediator-mapped database and genomic data are stored in an external database (e.g. BASE) which may not be mapped.) This is especially true if anonymization is occurring, as there will be in principle no way to manually check that the identifiers consistency is preserved.

2. The creation of mediator queries with the query builder may not be easy. The meaning of the various fields one has to choose from to create the query is not obvious. For instance, in the example of the MCMP/Hokkaido scenario (which is a trivial example of clinical database with only a couple of data in a few CRFs), there are four fields which are related to the sex of the patient that we have to choose from [from my memory, as the service is down at the time of writing]. It is not obvious to guess which one will return the data useful for the analysis. I guess more complex queries will be even more difficult to create (e.g. the follow-up information for a patient can be spread across many CRFs, and this information has to be aggregated at some point).

Ultimately, I think biostatisticians will likely create a single "mega-query" to retrieve all useful demographics/clinical information from the database, which may create performance issues.

3. In the creation of a data mining workflow, biostatisticians are working on a trial and error basis, e.g. by extending the analysis steps in a GridR component. The date required in data mining may/will thus evolve in time, meaning that the mediator queries will have to be adapted. The issue here is that the column identifiers associated to some given data may be altered if the query is modified, which may require modifying all GridR scripts retrieving data from the query. (The alternatives are a) to create queries which are specific to each component, or b) to have an understanding of how column labels are generated by the mediator to be able to adapt the queries in a way that they remain compatible with previous ones.

Suitability of the ACGT approach in this particular case

The inclusion of ArrayExpress in the ACGT data access platform presented some issues not encountered with previous data sources. The public nature of the repository with a proprietary data model, and the nature of genomic data led to some peculiarities that forced to adapt previous approaches for data access policies. First of all, the newly developed wrapper had to be adapted to the programmatic access interface provided by ArrayExpress, which offered a very low level of expressiveness possibilities. An ad-hoc RDF-base schema resembling the MAGE-OM object model had to be created as well. Second, the nature itself of the types of queries that end-users perform against combinations of clinical and genomic repositories implied new difficulties. When carrying out data analysis for both clinical and genomic data, the latter being stored in ArrayExpress, values extracted from the clinical database are employed to query the genomic data resource (in this particular case, accession numbers were first obtained from the SIOP clinical database, and those values were employed to retrieve the desired data from ArrayExpress). This poses a new difficulty in the mediation process. Our approach, as any other existing mediation approach, based its functioning on a global

schema to which the underlying database schemas were mapped. This allowed the Semantic Mediator to easily circumvent the existing semantic heterogeneities. The translation process produced the subqueries that should be submitted to each of the databases. This approach no more worked for the new situation, since it produced subqueries that attempted to retrieve the complete ArrayExpress database (filtering by accession numbers would be performed later, when those values would have been obtained from the clinical data resource). Of course this works in theory, but not in practice. It is technically unviable to gather so much data for a single query, mostly considering that end users demand a minimum responsiveness level from out mediation software. The solution was not to drop the approach nor to modify it, but rather expand it. The ontology-based approach for integrating disparate data sources is maintained, as this was not the source of the technical difficulties. Instead, the capability of the mappings was increased, leading to the ability of the mediator to carry out internal workflows to reduce to a minimum the data flow. The mapping with ArrayExpress is defined in a way in which a subquery for it is not produced until all possible information about accession numbers from the rest of databases implicated in the query is not available. This way, the query for ArrayExpress is more specific and it can be accomplished in a reasonable time.

Without a mediation system offering access to both clinical and genomic databases in a homogeneous manner, end users were forced to carry out manual workflows, or make use of existing workflow tools to retrieve all necessary data for their analysis. By adding support for ArrayExpress in the ACGT Semantic Mediator, researchers can bypass this bottleneck and reduce their work-load when dealing with the data access layer in the ACGT platform. More tests need to be carried out in order to ensure this approach offers a sound solution to the problem of clinical and genomic data integration. So far, the most important technological challenges seem to be solved.

7 Bibliography

[Petrik et al. 2006] Petrik V, Loosemore A, Howe FA, Bell BA, Papadopoulos MC (2006) OMICS and brain tumour biomarkers. Br J Neurosurg 20(5):275-280

[Ippolito et al. 2005] Ippolito JE, Xu J, Jain S, Moulder K, Mennerick S, Crowley JR, Townsend RR, Gordon JI (2005) An integrated functional genomics and metabolomics approach for defining poor prognosis in human neuroendocrine cancers. Proc Natl Acad Sci USA 102(28):9901-9906

[Li et al. 2005] Li W, Kessler P, Williams BR. Transcript profiling of Wilms tumors reveals connections to kidney morphogenesis and expression patterns associated with anaplasia. Oncogene 2005, 24, 457-468.

[Zirn et al. 2006] Zirn B, Hartmann O, Samans B, et al. Expression profiling of Wilms tumors reveals new candidate genes for different clinical parameters. Int J Cancer 2006, 118, 1954-1962.

[Wang et al. 2005] Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005, 365, 671-679

[Brazma 2001] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 2001 December;29(4):365-371.

[Spellman 2002] Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, et al. Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol. 2002 August;3(9).

[Sarkans 2005] Sarkans U, Parkinson H, Lara GG, Oezcimen A, Sharma A, Abeygunawardena N, et al. The ArrayExpress gene expression database: a software engineering and implementation perspective. Bioinformatics. 2005 April;21(8):1495-1501.

Appendix 1 – D7.8 Survey

The D7.8 Survey was designed to gather the experiences and general impressions of all the participants in the different scenarios developed for ACGT so far. The answers are reflected in different parts of this document.

D7.8 Survey

This survey is aimed at people working in the ACGT project, and intends to estimate the necessary effort when including a new clinical trial in the ACGT Platform. Given the complexity of the platform, many internal components are involved in the process of aggregating a new trial in the data access layer. The imaginary scenario is the following: a new trial is to be connected to the ACGT technological platform. The questions below pretend to estimate the effort to be carried out by different participants of this process since the creation of the trial to the point where its data is shared with existing trials for joint analysis.

Some answers might depend on some condition. For example: I will have and effort of 1MM to adapt my code if the trial is related to nephroblastoma, but 0MM otherwise. In situations like this, please briefly indicate and explain the most important or probable cases.

Please write your name and your institution:

- **Name:**
- **Institution:**

Please answer the following questions (some questions are only to be answered in case of a positive answer to a previous question).

1. **Are you a technical developer within the ACGT platform? (YES/NO) (If YES, please answer 1a, 1b and 1c, if not, skip to question 2)**

   1a. What is the name(s) of your module(s)/service(s)?

   1b. Does the inclusion of a new trial in the platform imply new development in your module (please specify what you have to do)? If yes, what would be the estimated MM?

   1c. Does the inclusion of a new trial imply the generation of data models (XML configuration files, etc)? If yes, what would be the estimated MM?

2. **Do you work in the development/maintenance/testing of the ACGT Master Ontology? (YES/NO) (If YES, please answer 2a and 2b, if not, skip to question 3)**

   2a. What is your task with the MO?

   2b. Does the inclusion of a new trial in the platform imply any effort within this task (please describe it shortly)? If yes, what would be the estimated MM?

3. **Are you a legal advisor of the ACGT project? (YES/NO) (If YES, please answer 3a and 3b, if not, skip to question 4)**

   3a. What is your task within ACGT?

   3b. Does the inclusion of a new trial in the platform imply any effort on your side (please describe it shortly)? If yes, what would be the estimated MM?

4. **Are you an end-user or tester of the ACGT platform? (YES/NO) (If YES, please answer 4a, 4b and 4c, if not, skip to question 5)**

   4a. What type of work do you carry out with ACGT?

   4b. Does the inclusion of a new trial in the ACGT platform affect your work? If yes, what would be the estimated MM?

   4c. Do you need specific training when a new CT is included in the platform? Are the written resources enough, or your training requires specific advice?

5. **When a new scenario is included in the platform, does your work imply to communicate/coordinate with representatives of other components or end users? Please summarize the related tasks.**

6. **Finally please identify, from your point of view, what are the main "bottlenecks" in the process of including new trials in the ACGT platform.**

Appendix 2 – Arrayexpress wrapper perform document

```
<?xml version="1.0" encoding="UTF-8"?>
  <perform
    xmlns="http://ogsadai.org.uk/namespaces/2005/10/types">
        <documentation>
      Simple query across all tables in the database.
    </documentation>
        <launchQuery name="myActivityInstance">
      <query>
    PREFIX p0:&lt;http://www.ifomis.org/acgt/1.0#&gt;
          SELECT  ?patient_1 ?humanbeing_1
            WHERE {
            #   SPARQL help:   Annotation: [Patient, NameString]
          ?patient_1 p0:roleOf ?humanbeing_1. ?humanbeing_1 p0:hasGender
        ?gender_1. ?patient_1 a p0:Patient. ?humanbeing_1 a p0:HumanBeing. ?gender_1 a
        p0:Male. }      </query>
              <type>CSV</type>
        <SemanticMediatorOutput name="sparqlResults"/>
    </launchQuery>


  </perform>
```

Appendix 3 – SIOP+Arrayexpress integration mapping file

```
<?xml version="1.0"?>
  <!--<!DOCTYPE mapping SYSTEM "mapping.dtd">-->
  <mapping>
    <dbinfo>
      <dbid>
        AEWrapper
      </dbid>
      <wrapperurl>
        http://138.100.11.248:8080/wsrf/services/ogsadai/DataService
      </wrapperurl>
      <description>
        Mapping for the ArrayExpress DB
      </description>
    </dbinfo>

    <map>
      <entrydescription>
        Gender of Patient - Male
      </entrydescription>
      <path_list>
        <src_paths>
          <path>
            <int_entity composingid="Patient">
              http://www.ifomis.org/acgt/1.0#Patient
            </int_entity>
            <rest>
                    <int_link>
              http://www.ifomis.org/acgt/1.0#roleOf
            </int_link>
            <int_entity composingid="Gender">
              http://www.ifomis.org/acgt/1.0#HumanBeing
            </int_entity>                    <int_link>
              http://www.ifomis.org/acgt/1.0#hasGender
            </int_link>
```

```
                <int_entity>
                  http://www.ifomis.org/acgt/1.0#Male
                </int_entity>
              </rest>
            </path>

          </src_paths>
          <target_paths>
            <path>
              <int_entity composingid="Patient">
                http://miNamespace#org.biomage.BioMaterial.Treatment
              </int_entity>
              <rest>                    <int_link>
                   http://miNamespace#Action
              </int_link>
              <int_entity composingid="Gender">
                http://miNamespace#org.biomage.Description.OntologyEntry
              </int_entity>
              </rest>
            </path>
          </target_paths>
        </path_list>
      </map>

  </mapping>
```