



D5.9 Report concerning lessons learnt and synergies with external initiatives

D5.8 Investigation of providing support to users concerning the exploration of available data sets

Project Number: FP6-2005-IST-026996

Deliverable id: D 5.8 & D5.9

Deliverable name: Lessons learnt and investigation on providing support for the exploration of available datasets

Submission Date: 08/09/2010



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:	D 5.8 & D5.9
Document name:	Lessons learnt and investigation on providing support for the exploration of available datasets
Document type (PU, INT, RE)	PU
Version:	1.0
Submission date:	08/09/2010
Editor: Organisation: Email:	Anca Bucur, Jasper van Leeuwen Philips Research {anca.bucur,jasper.van.leeuwen}@philips.com

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

ABSTRACT:

In this deliverable we discuss the main ACGT results with respect to providing uniform access to relevant heterogeneous data sources and identify the main learning points that will drive our future work. The deliverable also describes state of the art solutions for enabling the users to explore the available datasets, a need identified as relevant in ACGT. Part of this deliverable, we also look at possible synergies with relevant external initiatives and summarize the results of the study concerning the use of ACGT outcomes to support a large and innovative research program of the Breast International Group, NeoBIG. Based on the requirements of the NeoBIG program we have evaluated the benefits and the issues related to the ACGT approach, the technical solutions and the expertise that we could bring to support NeoBIG and the new research work that is required. The future research work with focus on the NeoBIG program will be part of the INTEGRATE project which is briefly described at the end of this document.

KEYWORD LIST: clinical trials, databases, data integration, data access services

MODIFICATION CONTROL			
Version	Date	Status	Author
0.10	20/04/2010	Draft	J. van Leeuwen
0.5	12/08/2010	Draft	A. Bucur
0.6	01/09/2010	Draft	J. van Leeuwen
1.0	08/09/2010	Final	J. van Leeuwen

List of Contributors

- Anca Bucur, Philips Research
- Jasper van Leeuwen, Philips Research

Contents

1	EXECUTIVE SUMMARY	6
2	INTRODUCTION	7
2.1	SCOPE	7
2.2	STRUCTURE.....	7
3	DATA ACCESS SERVICES	8
3.1	INTRODUCTION	8
3.1.1	<i>Technical section</i>	10
4	LESSONS LEARNT	14
4.1	EXPLORING DATA RESOURCES IN THE ACGT PLATFORM	14
4.1.1	<i>Relevant state of the art</i>	15
4.2	REPORTED ISSUES.....	22
4.2.1	<i>Query limitations of the common query language</i>	22
4.2.2	<i>Structured versus unstructured data sources</i>	23
4.3	USAGE FOR THE NEOBIG TRIALS	23
4.4	FROM THE NEOBIG REQUIREMENTS TO INTEGRATE	26
4.5	OTHER RELEVANT INITIATIVES – CABIG.....	29
	<i>References</i>	32
	<i>Appendix 1 - Abbreviations and acronyms</i>	34

1 Executive summary

In this deliverable we discuss the main ACGT results with respect to providing uniform access to relevant heterogeneous data sources and identify the main learning points that will drive our future work. The deliverable also describes state of the art solutions for enabling the users to explore the available datasets, a need identified as relevant in ACGT. Part of this deliverable, we also look at possible synergies with relevant external initiatives and summarize the results of the study concerning the use of ACGT outcomes to support a large and innovative research program of the Breast International Group, NeoBIG. Based on the requirements of the NeoBIG program we have evaluated the benefits and the issues related to the ACGT approach, the technical solutions and the expertise that we could bring to support NeoBIG and the new research work that is required. The future research work with focus on the NeoBIG program will be part of the INTEGRATE project which is briefly described at the end of this document.

While uniform data access to external heterogeneous resources is important for NeoBIG, we have also identified a need to consolidate data into large coherent datasets in a common repository, based on shared standards, methodologies and terminologies/ontologies. The ACGT solutions with respect to security and data privacy can be further used and extended to support NeoBIG. Other lessons learnt refer to building loosely coupled flexible services without a heavy middleware to be maintained, and the need to address sustainability of results early in the development process to be able to build solutions that can be used by large communities of users far beyond the end of the project.

2 Introduction

ACGT is a European Union co-funded project aiming at developing open-source, semantic and grid-based technologies in support of post genomic clinical trials in cancer research.

Advances in post genomic research have created significant opportunities for offering personalized treatment and better health care services to the population at large. At the same time clinical trials have become a bottleneck in terms of complexity, effectiveness and, in their present form, fitness for purpose. On the other hand, in the realm of information technologies advances in semantic technologies and grid computing have reached a stage where multi-dimensional applications requiring the combination of heterogeneous data and software resources can be realistically tackled.

In this light the ACGT platform is being developed, offering a unified technological infrastructure which will facilitate seamless and secure access, and analysis of multi-level clinical and genomic data enriched with high-performing knowledge discovery operations and services in support of multi-centric, post-genomic clinical trials.

2.1 Scope

Clinico genomic clinical trials on cancer typically collect a variety of different data types. Within the ACGT platform, these various data resources are made available (such as Case Report Form (CRF) databases, imaging data (e.g. MRI, PET/CT scans) and microarray data). In addition, the platform allows users to dynamically integrate their own relational databases into the platform. Syntactic and semantic integrations needs to take place in order to provide seamless data access. Syntactic data integration handles differences in the formats and mechanisms of data access, whereas semantic integration deals with the meaning of information; it must handle the fact that information can be represented in different ways, using different terms and identifiers.

Now the ACGT project is concluding, we look back at the work that has been done relating data integration and carefully assess it. We don't only look at the data access services, but also the relevant related work such as the requirements analysis and consolidation of the NeoBIG scenario.

2.2 Structure

In this document, the data access services are introduced in Chapter 3. Chapter 4 describes the lessons learnt, issues related to the ACGT approach, possible future directions, and reuse and extension of the work done.

3 Data Access Services

3.1 Introduction

One of the main challenges in carrying out post-genomic research is to efficiently have access to all relevant data. In the context of clinical trials, this data is typically scattered geographically and resides in a variety of different systems. The data typically comprises clinical data collected on Case Report Forms (e.g. symptoms, histology, administered treatment, treatment response), imaging data (e.g. X-Ray, CT, MR, Ultrasound), genomic data (e.g. microarray data), pathology data and other lab data. Next to that there are many public biomedical databases that are relevant. These store information about gene and protein sequences, pathways, genomic variation, microarray experiments, medical literature, tumour antigens, protein domains, metabolites, etc.

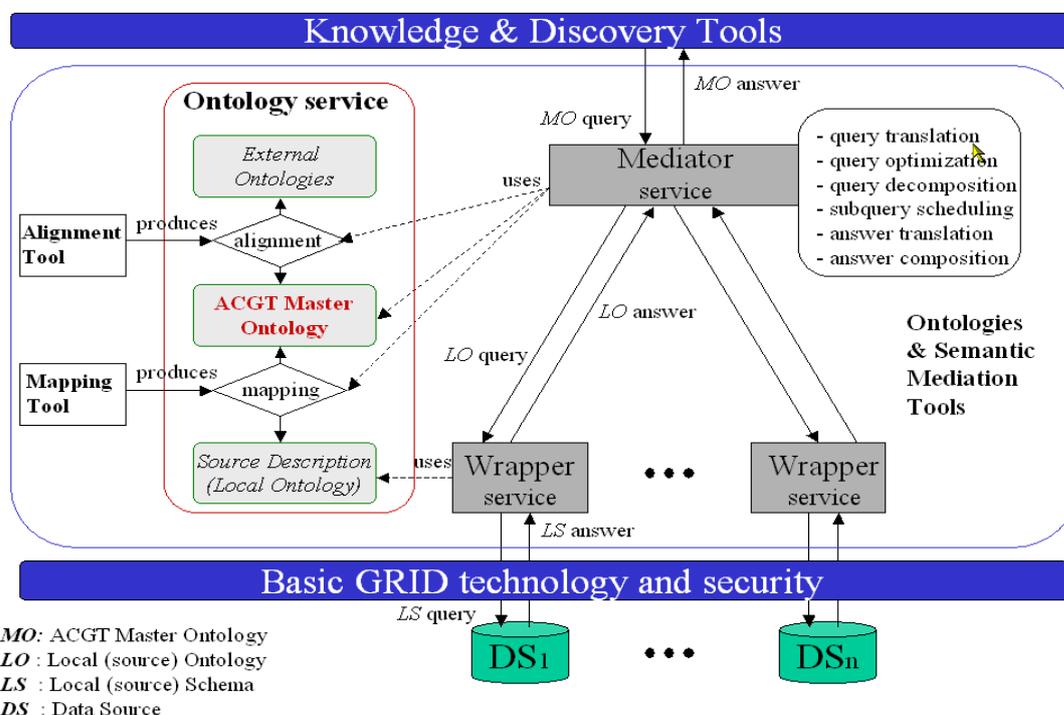


Figure 1 - ACGT information architecture

In order to overcome this challenge, the project has proposed the ACGT information architecture as depicted in Figure 1. The figure shows the ACGT information architecture as proposed at the outset of the project. It illustrates the envisaged role of the data access services, referred to in the figure as Wrapper services.

The end-user can interact with the ACGT platform in a variety of ways; one example is using the workflow environment in the portal. When the user wants to query data which is made available through the ACGT platform, the query is expressed in terms of the ACGT Master Ontology [7]. The ACGT Master Ontology (MO) is an ontology covering the domain of post-genomic clinical trials. It is a pragmatically constructed ontology, using the TOP trial [11] and Nephroblastoma trial [12] to guide its construction. The ACGT MO is an example of an application ontology as it spans

(parts of) multiple domains, aggregating various aspects of clinical trial management, cancer research and clinical care.

The Mediator service uses the ACGT MO to provide a virtual view on all data. Clients of the mediator do not have to know about location, schemas, or access methods of the underlying data sources. The mediator accepts queries expressed in the ACGT MO. It translates the query and decomposes it into one or more sub-queries that it issues to the data access services. These queries are expressed in the ontology of the underlying data sources, i.e. their local ontology. Thus, the mediator resolves semantic heterogeneities.

The data access services receive the queries from the mediator, transform these to the format of the underlying data source, and translate the results back to the local ontology schema. Thus, the role of the data access services is to provide homogeneous access to heterogeneous data sources, from the syntactic point of view. They hide differences in the access interface, the query language, the data format, etc.

We shortly introduce the data access services for background information. An extensive description of the ACGT the data access services can be found in [6]. The main functionality provided by the data access services is:

- **Support of a uniform mechanism for querying data.** Data can be queried using SPARQL, thus hiding the different query mechanisms provided by the underlying databases.
- **Export of the schema of data resources.** An RDF Schema of a data resource is available on demand. This schema is, for example, used by the semantic mapping editor to provide the mapping to the ACGT Master Ontology.
- **Retrieval and delivery of files.** Large files in standardised formats that do not need to be processed by the semantic mediator can be retrieved by ID, and delivered to specified locations (amongst others DMS). Files can optionally be compressed and combined into a single archive to speed-up delivery.
- **Credential-based authentication and authorization.** The ACGT data access services are fully integrated into the ACGT security infrastructure to control access to the data.
- **Dynamic deployment of new data resources.** To speed-up integration of new data resources, integration into the ACGT infrastructure can be carried out by the owner of the database using the ACGT web portal.

The data access services provision access to data for the ACGT platform. Currently, the data access services provide access to three main types of data; relational data, microarray data and DICOM image data. These types of data are typically used in clinical trials.

Relational data is used in multiple situations. Firstly, important clinical trial data is stored in these databases, for example data collected using CRF forms. Secondly, a variety of other data – for instance data used in the analysis of clinical trials – can be available in a relational database (e.g. the Gene Ontology - a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species – is available for download as MySQL database). And thirdly, relational databases can also be used to make data available that may not yet be stored in a relational database, but that can be mapped to the relational data

model. This holds for data collected in files of various formats, such as Excel files, plain text files (in CSV format for example), XML files, etc.

The data access services also provision access to microarray data stored in BASE (see [3]), as microarray data is often collected in post genomic clinical trials. BASE is an Open Source microarray database that is being developed by Lund University, one of the partners in ACGT. There are various advantages of using BASE as storing and analysis suite for microarray data; Base is plugin-enabled: Users can contribute analysis tools using a plug-in API, it supports many array formats, and custom formats can be used using a customizable import plug-in.

In addition, the data access services provisions access to DICOM image data. Imaging data (such as X-Ray, CT, MR, and Ultrasound) is routinely collected in clinical practice for diagnostic purposes and to assess response to treatment. Derived measures such as tumour size are typically collected on the CRF forms. Image data can also be used directly for simulation purposes. Imaging data is typically stored in a PACS (Picture Archiving and Communication System), which can be access via standardized DICOM interfaces. The data access services takes advantage of the DICOM interfaces, avoiding tight coupling with specific PACS products.

Another use case besides accessing/querying available data sources within the ACGT platform is the introduction of new data sources. This can be either new permanent data sources (such as data sources coming from a new trial) or temporary data sources (which are often used to assist in analysis). For permanent resources it pays off to do a complete semantic integration. Most of the time, the new data source will be a relational database. The Data Access Services support the semantic integration by generating a local ontology of the database, exported as RDFS scheme. This scheme is used by the Mapping Tool to generate the mapping with the ACGT MO. Optima (see [16]) however uses the ACGT MO to generate the scheme of the database. Therefore, the mapping with the ACGT MO can be generated completely automated. The Data Access Services also allow the user to dynamically add temporary databases to the ACGT platform. Temporary databases are not fully semantically integrated into the ACGT platform as the user typically owns the database (and is very familiar with it) and accesses it using the local ontology. This feature is very useful to quickly incorporate a (local) database to assist in analysis.

3.1.1 Technical section

Technically, the data access services are based on the OGSA-DAI framework. OGSA-DAI is a web services framework for providing data access (see [2]) in a document-style manner. Its activity framework enables efficient and flexible service invocation. The data access services are made compliant with the ACGT security infrastructure, using credential-based authentication and authorization carried out by the Grid Authorization Service as a Policy Decision Point.

The users of the data access services interact with the data access services in a document-style manner. A *perform document* describes the actions that a data service resource should perform on behalf of the client, whereas a *response document* describes the status of execution of a perform document and may contain result data, such as the results from a database query. Each action in a perform document is known as an *activity*, and can have (named) input- and output streams.

Of particular interest is the activity that provides SPARQL functionality. This activity is provided for the relation data resources and the DICOM image resources.

SPARQL (see [19]) is a query language for RDF (see [18]). RDF is a directed, labelled graph data format intended for representing information on the Web and is heavily used in semantic web technology as general method for conceptual description or modelling of information.

3.1.1.1 Relational data access services

The data access services provide a SPARQL activity for relational data. Internally, the relational data access provider uses the D2RQ platform ([4]) to enable an RDF-view on relational databases and to perform the SPARQL to SQL query translation. D2RQ is a declarative language to describe mappings between relational database schemata and OWL/RDFS ontologies. The mappings allow RDF applications to access the content of huge, non-RDF databases using Semantic Web query languages like SPARQL. The perform document contains the SPARQL query, the response document contains the RDF result triples (encoded into the format requested, e.g. XML or csv).

The D2RQ platform ([5]) is able to analyse a relational database schema of an existing database and generate a “default” mapping file. This mapping file is used to translate queries and/or the database content. An ontology is mapped to a database schema using *ClassMaps* and *PropertyBridges*. The central object within D2RQ and also the object to start with when writing a new D2RQ map is the *ClassMap*. A *ClassMap* represents a class or a group of similar classes of the ontology. A *ClassMap* specifies how instances of the class are identified. It has a set of *PropertyBridges*, which specify how the properties of an instance are created. In the default mapping, a table in an entity-relationship-model is mapped onto a *ClassMap*, and an attribute is mapped onto a *PropertyBridge*. Thus, D2RQ infers types of classes and properties and allows the user to provide labels, comments and additional properties. D2R Server also serves data for the associated vocabularies.

A user may introduce a new data resource within the ACGT platform. This for instance can be permanent data resources such as clinical trial databases or temporary relational databases to assist with a particular analysis. The data access service provides (amongst others) an RDF Schema of the data resource and a SPARQL interface to query the database resource. In case of a permanent data resource it pays off to create a semantic mapping for the semantic mediator, such that the data resources can be queried using the ACGT Master Ontology.

3.1.1.2 DICOM data access services

The DICOM data access services provide two main functions: querying of the image metadata (by means of SPARQL) and retrieval of specific images.

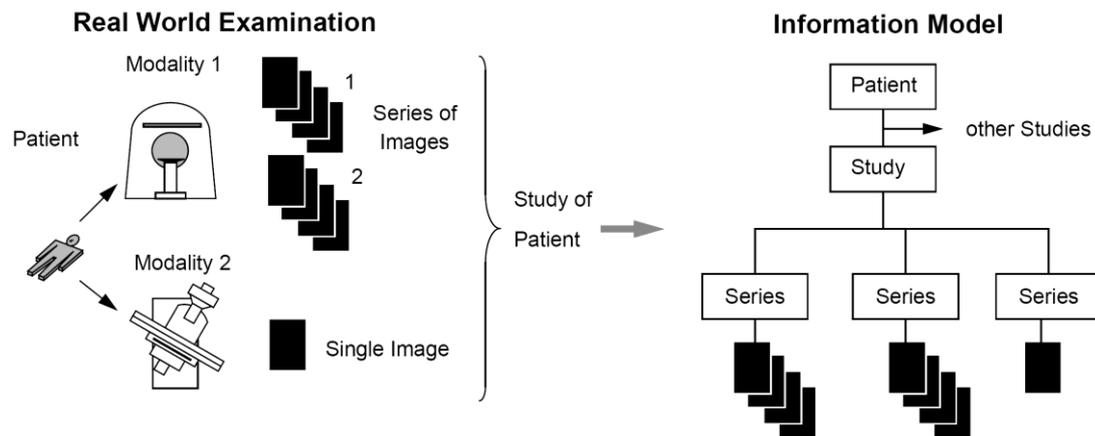


Figure 2 - The DICOM information model for examinations and its mapping to the real world.

The model that is queried closely follows the information model as defined within DICOM (see Figure 2). There are three information models that can be used for querying: Patient Root, Study Root, Patient/Study Only. The Patient Root Q/R information model contains four levels: Patient, Study, Series and Image. The Study Root Q/R information model is similar, except that the top level is the Study level. Attributes of patients are considered to be attributes of studies. The Patient/Study Only Q/R information model is also similar to the Patient Root model, except that it only supports the upper two levels. The models determine the type of queries that can be issued, but do not directly restrict what can be returned. For example, even though the Patient/Study Only model does not include images, images can still be retrieved by retrieving all images for a specific patient.

For each of the three Q/R models, and for each level in each model, the standard defines the attributes that can be searched for in the query. A DICOM server does not have to support all attributes. For each attribute it is stated whether they are required or optional. Most attributes are optional. For example, in the Patient Root model, the only two attributes that are required are Patient's Name and Patient ID. All other attributes at the patient level, such as Patient's Birth Date and Patient's Sex, are optional.

Because there is a difference in the expressivity between SPARQL and the way DICOM queries, the current implementation of the data access service imposes restrictions on the SPARQL queries it accepts:

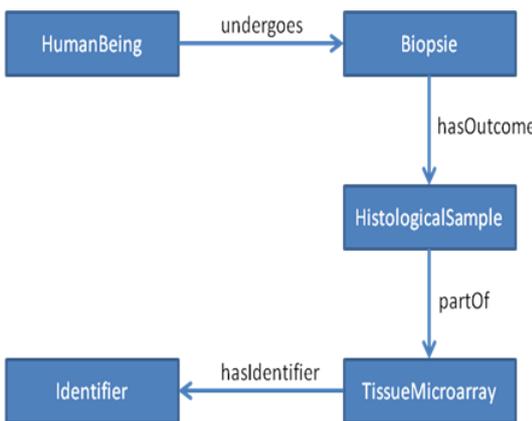
- A variable cannot be used as an object in more than one triple.
- A predicate in a triple must be a URI. I.e. it cannot be a variable.
- Optional blocks can only use predicates for a given level in the DICOM information model if a triple in the main block uses a predicate for that level.
- All triples with predicates corresponding to the same level of the DICOM information model must share the same subject.
- At every block in the query, all triples in scope must form a single connected graph.
- Queries cannot use SPARQL's "ORDER BY" or "DISTINCT" modifiers in the SELECT statement.
- Queries cannot use the UNION construct.

In practice, these limitations did not hamper the users.

3.1.1.3 ACGT Master Ontology

Within the ACGT platform, queries to the semantic mediator are expressed in terms of the ACGT Master Ontology (MO). The ACGT MO models the domain of cancer research and management (in a computationally tractable manner). The ACGT MO is based on a variety of sources (see [9]), most notably the Case Report Forms for clinical trials as provided by ACGT's clinical partners.

An example SPARQL query expressed in the master ontology, "retrieve the identifiers of the microarray files for each patient":

Query:	Diagram:
<pre> SELECT ?patient ?arrayID { ?patient a acgt:HumanBeing ; acgt:undergoes ?biopsy . ?biopsy a acgt:Biopsy ; acgt:hasOutcome ?tissueSample . ?tissueSample a acgt:HistologicalSample ; acgt:partOf ?microarray . ?microarray a acgt:TissueMicroarray ; acgt:hasIdentifier ?identifier ; ?identifier a acgt:Identifier ; acgt:hasStringValue ?arrayID . } </pre>	 <pre> graph TD HumanBeing[HumanBeing] -- undergoes --> Biopsie[Biopsie] Biopsie -- hasOutcome --> HistologicalSample[HistologicalSample] HistologicalSample -- partOf --> TissueMicroarray[TissueMicroarray] TissueMicroarray -- hasIdentifier --> Identifier[Identifier] </pre>

When integrating new data sources into the ACGT platform, it is possible that part of the domain is not yet covered by the ACGT MO. For this purpose, there are procedures defined (see [1]) to submit change requests to the ACGT MO. To facilitate this, a workflow and communication system - The Ontology Submission tool - has been created that gathers all the change requests regarding the content of the ACGT MO, feeds them to the ontology experts in a manageable way, keeps the version history of the ACGT MO, and automates the communication back to the interested parties of any changes taken place.

4 Lessons Learnt

4.1 Exploring data resources in the ACGT platform

Users can query a data resource and obtain its RDF schema when it is available within the ACGT platform. If the data resource is fully integrated into the ACGT platform, the queries are expressed in terms of the ACGT Master Ontology, otherwise the query are expressed in terms of the data resources local ontology.

The SPARQL query interface has a number of observed drawbacks. The query resolution is based on graph pattern matching to produce a solution sequence, where each solution has a set of bindings of variables to RDF terms. The soundness of the query graph is however not verified. For instance if the query graph contains a misspelled RDF term, it is likely that the solution sequence will be empty. This has the unfortunate property that whenever an empty solution sequence is returned to a SPARQL query, this is due to either:

- the query was sound and no solutions exist in the data resource, or
- the query was not sound (but solutions may exist to the intended query).

We consider this a major drawback of the provided SPARQL interface.

When first encountering a (potentially) useful data resource, users often try to get to grips with the data resource by exploring it. This is also the case when a semantic mapping with the ACGT Master Ontology has to be made for the semantic mediator in order to semantically integrate a new data resource into the ACGT platform. Depending on the layout of the underlying data resource, this can be quite cumbersome. This experience was described in [6], where the same dataset (for the Wilms tumor analysis scenario) was stored in two different database layouts. The first layout was the original SIOP database, which closely followed the structure of the actual CRF's as used in the clinical trial. There are specific tables corresponding to each (part of) a CRF, with columns mapping to specific fields on the CRF. The second layout was according the Obtima database scheme. The Obtima scheme is a generic clinical trial scheme, almost similar to a meta-schema. For the uninitiated user, the generic layout is impossible to understand on its own. There are several reasons for this. The structure of the data can be concluded directly from the ontology (RFDS) in the context of the specific trial for the first layout, where this is not the case for the generic layout. To understand the setup of the specific clinical trial in the generic layout however, one has to descend to the instance level of the data. The cause of this problem is the lack of a (preferably standardized) description of the content of the data resource aimed at the (human) user (as opposed to computer clients). In addition, there is currently no tool available to help to browse conveniently through the data contained in a data resource.

Another (related) issue is possibly the generality of the ACGT MO versus the specificity of the (legacy) database. The database (scheme) is developed with a specific concrete goal in mind, while the ACGT MO intends to be very generic in an attempt to represent the domain of cancer research and management. This has an interesting effect when trying to explore a data resource by formulating SPARQL queries using the ACGT MO as a guide. Because the ontology captures such a wide

domain it is very easy to specify a constraint on a term that is not used in the actual data resource, resulting in an empty solution as response.

4.1.1 Relevant state of the art

In this chapter we review what the current state of the art has to offer to resolve the issues mentioned in the previous chapter.

4.1.1.1 /facet

In [14], a browser (names /facet) for heterogeneous semantic web repositories is proposed. The browser is a generic RDFS/OWL facet browser. A facet highlights one dimension of the dataset. The browser allows the user to navigate through the dataset and to set (combinations of) constraints on the facets of the dataset, resulting in a selection of the dataset. When the query becomes over-constrained (resulting in an empty selection), the constraint is removed.

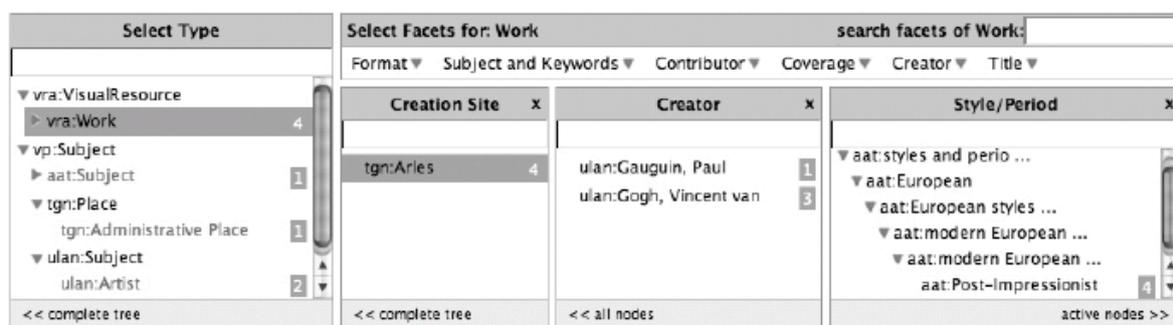


Figure 3 – slashfacet 15nrol15ncti specification

In their paper, an extensive example is given from the art domain. Figure 3 shows how a constraint is added to the constraint set. On the left side (“*Select Type*”), the types are shown (along with the amount of instances in the currently constrained data subset). On the right and upper side of the picture (“*Select facets for:...*”), those facets of the selected type are shown that have instance in the currently constrained data subset. Finally the bottom right side of the picture shows facets (e.g. *Creator*) and the associated constraints with the resulting dataset size. (e.g. constraining facet “*Creator*” to “*ulan:Gogh, Vincent van*” will result in a result dataset with 3 instances).

We consider this approach a promising way for resolving the issues as mentioned in the previous section.

4.1.1.2 Semantic Visualization of Patient Information

In [21] the mapping of patient data on relevant fragments of ontologies and inferred ontological structures is used as a basis for improved patient data visualization, comparison and analysis. Visualization is used for the comparative exploration of similar patients, which is considered a key requirement in CDS systems.

Patient data is complex and heterogeneous, stored in different formats and different structures, and bears different semantics. All these make comparison and analysis of clinical data a very challenging task.

The paper introduces several needs in healthcare whose resolution depends on better exploration of patient data:

- Clinical practice and research rely increasingly on analytic approaches to patient data.
- There is a growing need to store and organize all patient data that may contribute to a case evaluation or clinical study.
- Information is getting richer with more and different types of attributes being recorded, which makes the visualization of the overall picture increasingly difficult for the clinician or researcher.

The authors identify the presentation of similar patients as key requirement to enable medical decision support systems. To integrate medical data along multiple dimensions of heterogeneities is a significant challenge which needs to be solved to provide a coherent view of the collected clinical information.

The work aims to align patient data with relevant fragments of ontologies and to infer a more descriptive patient ontology as basis for improved patient data visualization, comparison and analysis. The authors assume that by mapping patient data attributes onto ontological concepts, the inherent structuring information of ontologies can be used for structuring and classifying the patient records. The external semantics allow for the patient record classification by a 16nrol16nc, from where the inferred hierarchy is directly fed into an appropriate visualization tool. Their tool based on the incorporation of medical ontological knowledge aims to contribute to improved and concise patient data visualization.

In this work, a hierarchical order of cases is achieved by building lexicographic hierarchies. These are created by imposing an a priori hierarchical structure on the attribute values which in turn induces a hierarchy on the dataset. By selecting a small number of attributes and by mapping those to concepts of an ontology, the structuring information of the ontology is used to obtain a hierarchical structure of the dataset. These lexicographic hierarchies where each cluster carries a label reflecting the corresponding semantics can be visualized. Browsing facilities over the set of all patients provide means for identifying similar patient records with respect to selected and relevant patient attributes.

The main use case described refers to brain tumors and incorporates more than 100 attributes of medical history and status data. The incorporation of semantics is achieved using inference – DL reasoning on an OWL DL view of patient data aligned with an OWL DL medical ontology.

The ontologies used are the Foundational Model of Anatomy to specify the tumor location (70000 distinct anatomical concepts and 1.5 mil. Relationships) and the WHO Classification of Tumors of the Nervous System for classification and grading. The WHO ontology establishes hierarchical structuring of histological typed tumors covering IHC, genetic profiles, epidemiology, clinical signs and symptoms, imaging, prognosis and predictive factors.

With this tool, clinicians can identify all patients with similar brain tumor locations, where the similarity measure reflects different level of detail in patient data description. According to their evaluation, clinicians appreciate to locate patients that are similar with respect to some selected classification axis for subsequently accessing their treatment process and progress and compare them.

4.1.1.3 Architecture for Semantic Navigation and Reasoning with Patient Data

In [13] an architecture enabling semantic navigation of patient records and reasoning with patient data is described. Their motivating use case is that the clinician believes that an extensive set of patient data would reveal subtle patterns if visualized in the right way. This work is closely linked to that described in the previous section and describes the architecture of the tool developed to enable better visualization of patient data records and the identification of similar cases by making use of ontologies and building lexicographic hierarchies.

Medical ontologies are the standard means of recording and accessing conceptualized medical and biological knowledge. In the context of patient information, their application is primarily annotation of medical data. This paper proposes the visualization of instance data with the help of knowledge that is represented by the ontology.

Several key requirements for clinical decision support are identified and addressed in this work:

- Clearly arranged presentation of annotated patient data
- Presentation of similar patients with respect to the complex and heterogeneous patient data
- The discovery of patterns and dependencies in patient data.

It is argued that with traditional applications little or no help is given to interpretation because the semantics are implicit and therefore inaccessible. The FMA and the WHO ontologies are used to extract the relevant sets of concepts and their relationships. They identify manageable sizes of ontologies that are suited; the selection is manual and the segmentation of the ontology is described as a difficult task.

The WHO classification refers to the ICD-O code and includes a WHO grading scheme that is used for predicting response to therapy and outcome. They use the WHO classification's inherent hierarchical structuring for hierarchically representing patient data.

The architecture includes the following main components:

- An ontology manager that can integrate the different knowledge components
- A mapping mechanism between the database schema and the OWL-DL A-box
- A reasoner which answers DL queries
- An optional Ontology Transformation Component is used in the hierarchical classification use case to establish the set of queries or labels
- The Interpretation & Visualization Component maps the inferred classification from OWL to the appropriate representation of the user interface. This component can also add attributes from the database which were not part of the reasoning process.

The system combines reasoning about data using external knowledge with advanced visualization that makes use of the inferred structure. The paediatric brain cancer data that was used to build this system was acquired in the EU-funded Health-e-Child project.

Several challenges have been defined related to this work. While a number of approaches to visualizing hierarchies exist, their power can only be fully exploited if the clinicians are trained to define the visualization which suits their interest and then navigate it. Inferring the hierarchy of the data based on its semantics requires suitable ontologies and alignment with them. However, a serious bottleneck is the 18nrol18nc whose performance scales very badly with the size of the ontology; this requires a proper extraction of the relevant fragments, however the ontology segmentation is also difficult, time consuming, manual, and can only be done off-line.

4.1.1.4 Ontology-Driven Automated Generation of Data Entry Interfaces to Databases

In [8] an approach for the semi-automatic generation of data entry interfaces to databases is described. An existing domain ontology is mapped to a system domain model, which can be specialised by domain experts to suit the data entry needs of their projects. The approach is applied in biological taxonomy, a domain in which capturing semantically-well defined data is essential. This branch of biology classifies the organisms into an ordered hierarchical system of groups reflecting their natural relationship and similarities.

The challenge addressed in this paper is the design of data entry interfaces to databases that enable the capture of high quality data without overburdening the user. Most solutions to data entry can only constrain data entered to conform to the data type associated with the table attributes. In most databases it is hard to ensure consistent use or data quality especially in terms of the semantics of the attributes. In order to have meaningful data in long term and to achieve data integration across databases, it is important to capture the semantics of the data along with the actual data. The semi-automatic data entry interface generation tool presented in this paper aims to help the quality of data entry to databases. The system generates an interface that reflects the semantics of the data as captured in a domain ontology and improves the semantics and rigor of the recorded data whilst minimizing the burden of data capture.

The design of the system adopts a model-based approach that includes task, domain and presentation models. The data entry task model is encapsulated in the system and the only change allowed is the order in which the data entry task is completed.

Several domain models are used to represent domain knowledge. An abstract domain model is transformed by mapping to it an existing domain ontology into a concrete domain model. It is only necessary to perform this mapping once for a given domain conceptual model. This model is further specialised by an expert into a specialized domain model for a given project. Two presentation models, one for ontology presentation and one for data entry, are defined in the system.

Model-based automatic generation approaches have not been widely adopted and have been criticized for not being able to produce quality, appropriate interfaces. This approach however provides a modelling tool for domain experts to specialize the domain model, and not for interface developers. The experts' intervention can ensure that the generated interface is suitable for the task at hand and that is specialized for the project that uses it. The tool developed has limited the possible permutations of the interface by limiting the approach to a descriptive data entry task, which allows the presentation model to be more appropriate.

4.1.1.5 Semantic Representation and Querying of caBIG Data Services

In [17] a model enabling users to query an integrated view of caBIG data services at a conceptual semantic level is described. This is based on semCDI, a previously developed model that enables researchers to utilize a single conceptual representation of the data instead of the various distinct models defined by each of the underlying data services. semCDI is applied to generate an ontology view of the caBIG semantics, and the underlying sources are queried using SPARQL queries complemented with Horn rules.

This work is carried out in the context of cancer Biomedical Informatics Grid (caBIG) Initiative. Interoperability is addressed by caBIG using a design consisting of a syntactic layer and three semantic layers. Interface integration is handled at the syntactic layer. The first semantic layer is the controlled terminology layer and is maintained in the NCI Thesaurus, a reference terminology published by the Enterprise Vocabulary Service. It contains a list of all concepts that the caBIG semantic structure recognizes. In a second layer, each of these concepts is linked to one or more common data elements (CDEs). A CDE identifies a property that can be associated to a concept, and it assigns a value restriction or a value domain to that property. The third semantic layer is the domain model layer which is a collection of UML models of the caBIG compliant data services. These models are used to link the data source metadata to caBIG's concepts and CDEs, and are contained in the cancer Data Standards Repository (caDSR).

semCDI is a query formulation that defines an ontology view of caBIG semantics where:

- Terminology concepts and domain model classes are modelled as ontology classes
- Associations between domain model classes are represented as object properties
- Attributes encoded in CDEs are modelled as data type
- Data objects are modelled as OWL individual members of the corresponding domain model class.

semCDI then uses SPARQL to pose queries against this ontology view. Definite Horn rules are used to define a priori conditional statements that are not explicitly asserted by the ontology extracted from caBIG. These rules are defined outside the query to allow the use of the same rules with multiple queries independently.

The authors also report several grid-specific issues that were encountered and which they hope to be resolved by the maturing of the caBIG technologies:

- Many of the caBIG data sources had service outages ranging from few hours to few days.
- When queries involve large data sets, the time required to formulate an ontology or receive query results ranges from two minutes to twenty minutes.
- Several of the caDSR domain models contain internal inconsistencies.

Other recent work has focused within caBIG to develop an infrastructure of data identifiers that would uniquely identify concepts on the grid. The authors aim to

leverage on those efforts and use these identifiers as a standard Horn rule applied on the result sets, to reduce inconsistencies.

4.1.1.6 Three Decades of Data Integration – All Problems Solved?

In [20] an overview of approaches to address data integration is provided. The issues and the approaches described are seen from a database and architectural perspective. The authors argue that the most difficult integration problems are caused by semantic heterogeneity and that these problems are being addressed by focusing on applying explicit, formalized data semantics to provide semantics aware integration solutions.

The reason for integration is presented as twofold:

- First, given a set of existing information systems, an integrated view can be created to facilitate information access and reuse through a single information access point.
- Second, given a certain information need, data from different complementing information systems is to be combined to gain a more comprehensive basis to satisfy the need.

The aim of the integration of multiple information systems generally is to combine the selected systems to form a unified new whole and give users the impression of interacting with a single system. Therefore, users are provided with a homogeneous view of data that is physically distributed over heterogeneous sources. For this all data needs to be represented according to the same abstraction principles: unified data model and unified semantics. This task needs to include the resolution of schema and data conflicts regarding structure and semantics.

The authors believe that in general information systems are not designed for integration which makes this task difficult. While the goal of integration is always to provide a homogeneous, view on the data from different sources, the integration problems are various. The particular integration task may depend on several factors:

- The architectural view of an information systems
- The content and functionality of the component systems
- The kind of information that is managed by the component systems
- Requirements concerning the autonomy of the component systems
- Intended use of the integrated information system
- Performance requirements
- The available resources (time, money, etc.)

The integration approaches are distinguished according to the level of abstraction where integration is performed. Describing information systems through a layered architecture with several layers, the integration issue can be addressed at each of the layers. The layers defined are: users, user interface, application, middleware, and data management (including two sub-layers: data access and data storage).

Therefore, the following general approaches are available according to this classification:

- Manual integration, where the users directly interact with all relevant information systems and manually integrate data. Users need to deal with different user interfaces and query languages, and know the data location, the logical data representation and the data semantics.
- Common user interface: The user is provided with a common interface that gives a uniform look and feel, but data is still presented separately and the homogenization needs to be done by the users.
- Integration by applications: With this approach integration applications access the various data sources and provide integrated results to the user. This solution is practical for small scale integration, but does not scale well with the numbers of system interfaces and data formats that have to be homogenized.
- Integration by middleware: A layer of reusable functionality is provided which solves all dedicated aspects of integration. While applications are relieved of common integration functionality, integration effort is still required in each application.
- Uniform data access: A logical integration of data is accomplished at the data access level. Applications are provided with the unified view on the data, but this approach is time consuming as data access, homogenization, and integration have to be performed at runtime.
- Common data storage: Physical data integration is performed by transferring data to a new storage. Local sources can either be retired or remain operational. When local sources remain operational, periodical refreshing of the common data storage may be required.

In practice, integration solutions are built based on these approaches. Important examples include mediated query systems (uniform data access), portals (uniform data access), data warehouse (common data storage), operational data stores (common data storage), federated database systems (uniform data access), etc.

4.1.1.7 Ontology Visualization Methods – A Survey

In [15] several ontology visualization methods and also a number of visualization techniques used in other contexts that could be adapted for ontology visualization are presented.

The authors argue that visualization of ontologies is a difficult task as they are more than a hierarchy of concepts. They are enriched with role relations among concepts and each concept has various attributes related to it and instances that can range from one to thousands. Their survey aims to provide useful information for choosing an ontology visualization for a specific application, taking into account functional and non-functional requirements and tasks that are related to the specific application.

The methods can be grouped according to different characteristics of the presentation, interaction technique, functionality supported, or visualization dimensions. In this survey the methods were grouped in the following categories representing their visualization type:

- Indented list
- Node – link and tree
- Zoomable
- Space-filling
- Focus + context or distortion
- 3D information landscapes

Methods in each of these categories may have elements of other categories. This classification has been chosen because each category has characteristics that lead to advantages and down sides.

The authors conclude that there is not one specific method that is the most appropriate for all applications and that the best solution is to provide the users with several visualization methods and allow them to choose the one that best suits their current goals. Several existing tools, such as Protégé, include several visualization plugins to provide a combination of several visualization methods. However, when the application is best suited by choosing a single visualization of the ontology, the designer needs to make a choice based on several characteristics of the ontology, the user profile, the application, etc. This extensive survey attempts to provide guidance in such cases.

An important conclusion of the paper is that visualization also needs to be coupled with effective search tools or querying mechanisms. Browsing is not enough, especially in the case of large ontologies.

4.2 Reported Issues

Although the current data access services provide most of the functionality that is required, there are still some open issues that could be addressed to further improve the functionality of the data access services. These are discussed below.

4.2.1 Query limitations of the common query language

A drawback of using a common query language is that not all queries supported by the underlying data sources can be expressed in it. This means that when the data access services are used, certain queries cannot be carried out as efficiently as would be possible by querying the underlying data sources directly.

In the case of SPARQL, the lack of support for aggregation (averaging, summation, counting, etc) is most likely to become an issue in practise. We already established in the requirements document that users occasionally want to carry out queries that use aggregation (see e.g. Section 4.1.2 of D5.1). This on its own does not imply that aggregation needs to be supported by the data access services, as aggregation can be performed client-side. However, this approach has a performance penalty associated with it. Fortunately, the SPARQL community has recognized the need to support aggregation, as this is included in the new SPARQL 1.1 Query Language specification.

4.2.2 Structured versus unstructured data sources

The idea to much more integrate clinical practice and clinical research gets more and more traction. This will have large consequences such as a higher innovation pace in patient care, and improved patient safety (e.g. due to less duplication of patient data entry). The context of the ACGT project is the realm of Clinical Trials. Compared to clinical practise, this has an effect on the way data is recorded. In clinical practise, the primary and often only purpose for data recording is direct patient care (on a per-patient basis), and a lot of data is recorded in free text format. In ACGT, there is also the clinical trial aspect. Due to the clinical trial requirements (to allow for a correct and efficient analysis), the data collected is typically structured. When generalizing the ACGT effort to include data collected in clinical practise, it should be investigated how to integrate free text into the platform (this includes natural language processing, and semantic integration of the results). Especially in a European context, the language processing difficulties are greatly multiplied (compared to the US).

4.3 Usage for the NeoBIG trials

In deliverable 5.7 [6] an analysis was carried out, providing the initial requirements for the NeoBIG program and analysis for the data sharing platform. The NeoBIG program is a research program led by Breast International Group, and aims to accelerate drug and biomarker development in early breast cancer, recognizing that the current drug development process is suboptimal and aims to improve the results of clinical trials. A durable, multidimensional translational research structure supporting neo-adjuvant trials will be build, sharing strategies, expertise, technologies, methodologies and protocols. In addition this will provide a strong foundation for future adjuvant trials in breast cancer (and research in other cancers). The program should result in a lasting bioinformatics platform for collaboration between cancer research institutes in Europe.

A new model trial design (see [10]) is proposed to speed up the evaluation of potential new drugs and biomarkers. NeoBIG trials will use selected patient populations (based on molecular subtype) in a Neo-adjuvant setting, such that surrogate endpoints will enable quick go-no-go decisions. Trials will be based on an integrated biomarker program (using gene expression signatures, circulating tumour cells, proteomics, genetics, and functional imaging), and will contain a central pathology review; all resulting in a pipeline of targeted agents.

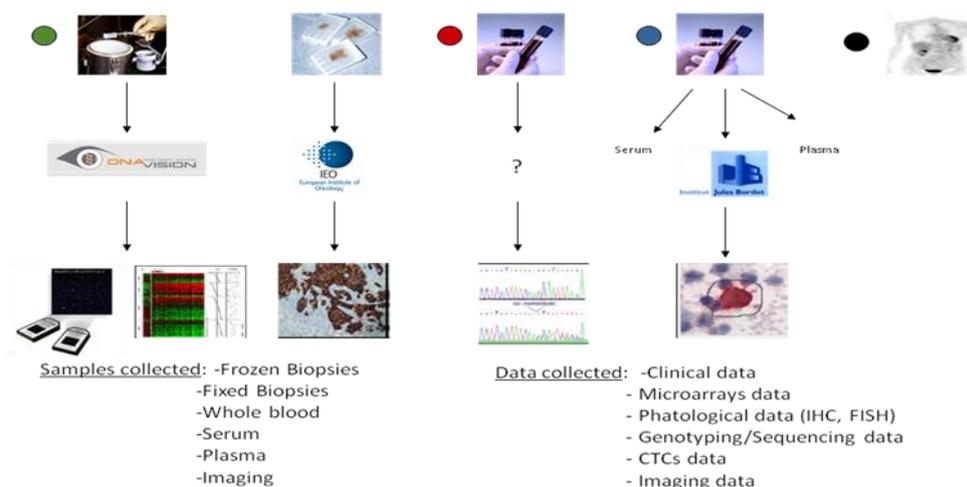


Figure 4 – Data collection

In order to exploit the clinical trials to the fullest, it is apparent that access to the data generated by a clinical trial cannot be restricted to a single individual or institution. Therefore, the NeoBIG project envisions a platform in which data gathered by the clinical trials can be easily accessed, but at the same time carefully controlled. In this document, the requirements for such a platform are laid out. This platform is envisioned to provide secure and controlled access to the data gathered in the clinical trials for various parties. The platform thus will contain a wide variety of data types (see Figure 4), ranging from the Case Report Forms collected in the clinical trials to the latest technologies in the genomic and proteomic fields (e.g. molecular data, microarrays, imaging, etc.). The platform should result in a lasting bioinformatics platform for collaboration between cancer research institutes in Europe and have a strong focus on interoperability.

In principle, the data should be gathered electronically. However, there should be accounted for that the infrastructure of the participating partners might not yet support this (thus an alternative path should be possible).

For the trials there will be different setups in different countries, but the repository could be managed and maintained by BrEAST. Especially in the UK or Germany there may be local repositories as well. The BrEAST data centre is the default data centre. They do the digitization of the CRFs. The access could be widened to the data storage.

Data will be maintained for at least 5 years and there will be around 1500 patients from the first five trials. But when the results are promising new trials may be initiated. The life of the repository should be significantly beyond the duration of the trials.

The data collected from the standard arms could be used to plan new research and to refine the results through follow up trials. This data should be stored and managed by the data sharing platform and the sharing of the data in a secure and privacy-preserving way among the members of the NeoBIG consortia and other authorized parties (potentially at a cost) should be supported.

The experimental arms of each trial should only be available to the parties participating in that trial, when they are authorized to access the data.

Currently the only communication requirement is that all centres have e-mail and fax access. eCRFs could be an interesting option, but costs are considered very high. There is a push from the pharma companies to introduce eCRFs. But the current priority of neoBIG is the gene expression and the clinical data.

To provide a platform that enables data sharing and collaboration between cancer research centres, NeoBIG requires a robust, secure IT solution that is compliant with a wide set of regulations and laws in the context of security, safety and privacy protection. The platform needs to be able to store, manage, and share the various types of data that will be generated by NeoBIG trials.

Security is an important aspect of the NeoBIG data sharing infrastructure. NeoBIG deals with personal data obtained from patients, whose privacy needs to be protected (both from an ethical and a legal perspective). Secondly, future prospective clinical trials with targeted therapies will require a system capable of dynamically setting up collaborations of organizations around specific data sets. Data shared within such a group needs to be well protected. Therefore, the NeoBIG data sharing platform needs to assure secure data sharing, such as authentication of users (secure logon), authorization (access control), encryption (to guarantee confidentiality), trust establishment, and Virtual Organization Management. Additionally, the interactions with the NeoBIG data sharing platform need to be fully audited to enable traceability.

Strong requirements on the data sharing platform are production-level reliability and availability and full maintenance. The data sharing platform will be used and needs to be available long beyond the end of the clinical trials, as the data is highly valuable for further research. Additionally, data interoperability and adherence to widely accepted international standards are important requirements which will enable the collaboration between BIG and other cancer organizations world-wide. In that context, well-known standards (HL7, DICOM, MIAME, MAGE, etc.) and terminologies (SNOMED, LOINC, etc.) are relevant, but also new standards emerging with the development and adoption by the US research community of relevant NeoBIG tools. As collaboration with the US cancer research community is desired and the US market is important for the pharma organizations participating in the NeoBIG trials, additional requirements need to be extracted from regulatory frameworks (such as FDA 21 CFR part 11) to which compliance needs to be assured.

We have concluded that there is a lot of ACGT expertise that could be used for the neoBIG data sharing platform, especially with respect to data storage, management and sharing, and with respect to privacy and security. On the other hand, due to the very strict requirements for a production-level system, with available documentation and user support, commercial deployment and long term maintenance, we have concluded together with the BIG that current ACGT prototype tools and services cannot be directly used for the neoBIG project. The same was preliminarily concluded about available caBIG tools and services. caBIG should still be evaluated as several standards emerging from that community should be taken into account in the development of the platform to enable interoperability and facilitate the collaboration between BIG and the US research community. We can conclude that although due to the prototype status of our solutions and to the fact that we are not able to provide commercial service level agreements, long term maintenance and some of the more focused requirements such as certification, our solutions cannot be directly used in the neoBIG scenarios, much of our expertise and ACGT work in

privacy, security and data access have proven highly relevant for the neoBIG data sharing platform.

4.4 From the neoBIG requirements to INTEGRATE

Part of the evaluation of the ACGT results, we have investigated how the expertise, and potentially also the tools, developed in ACGT could be used to support a large real-life multi-centric clinical trials program, such as NeoBIG, the new research program of the Breast International Group. To suit the ACGT scope, our focus was on the IT needs of the neoBIG research program, specifically with respect to secure privacy-preserving data management and sharing as these are issues at the core of ACGT. We have evaluated the suitability of our solutions by first collecting and analyzing the requirements of BIG concerning the data sharing platform needed to support their future clinical trials, and based on that briefly evaluating potential alternatives in which ACGT could support this program by making use of ACGT tools and infrastructure, but also of relevant expertise.

We have collected requirements by carrying out interviews and discussions with the main stakeholders of the data sharing platform, i.e. with representatives of BIG for the clinical aspects and with representatives of the Breast European Adjuvant Studies Team (BrEAST), the data centre of the BIG, for an IT perspective. Central to the discussions were requirements and scenarios concerning the building of a data sharing platform to support the NeoBIG program of the Breast International Group, that will be sustainable far beyond the neoBIG program and provide the clinical research community with common methodologies and standards, data models and consolidated datasets that could be used for further research.

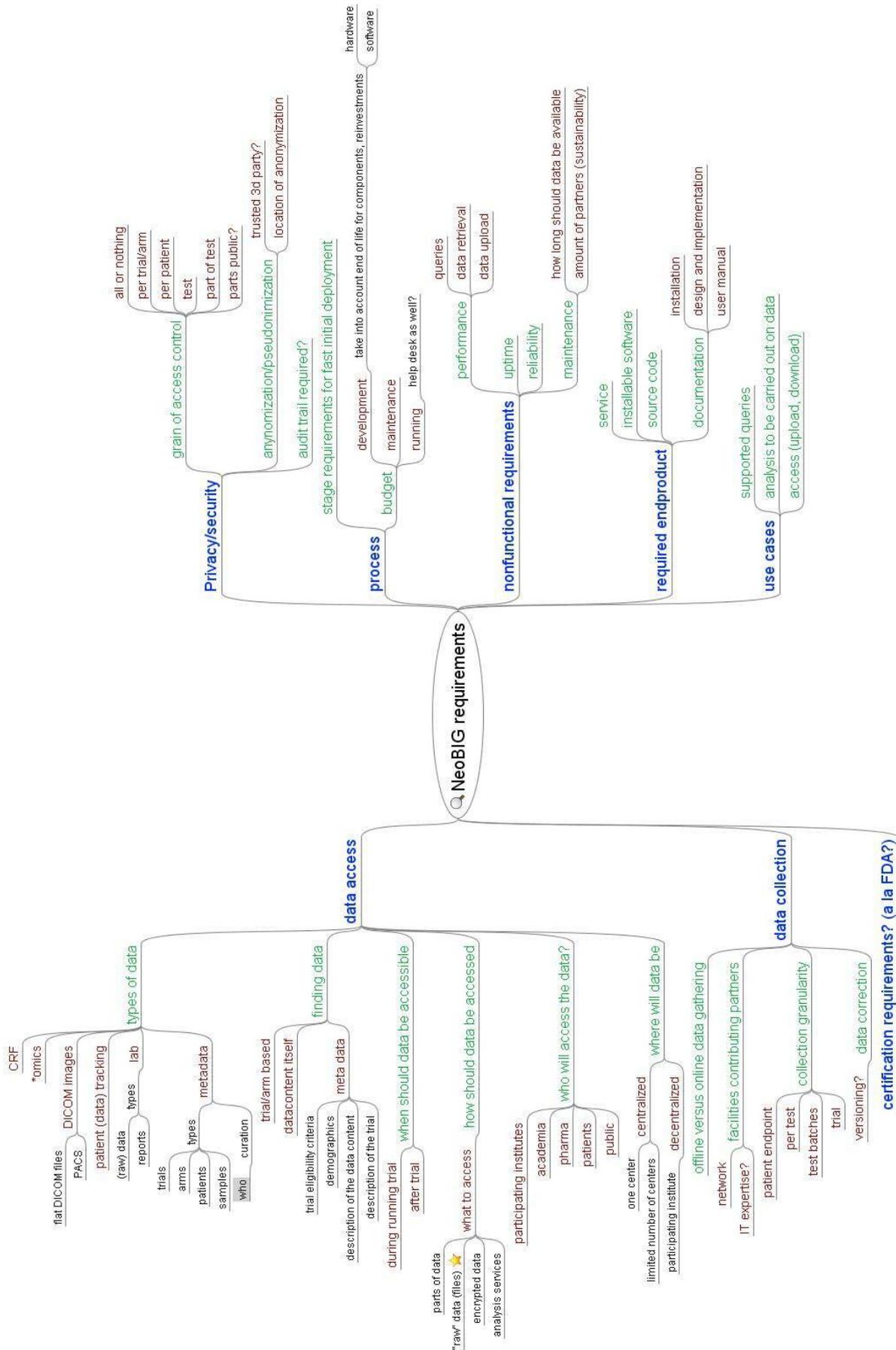


Figure 5 - Requirements of the NeoBIG program

During the study, we have concluded that while accessing external data out of heterogeneous repositories is highly relevant (therefore data access research carried out during ACGT could be used for NeoBIG), there is also very high value in supporting the neoBIG community to build comprehensive datasets including all the wealth of data collected in the neoBIG trials, and to provide infrastructure enabling large scale collaboration and sharing. The advantage of building such consolidated data sets under a single authority in charge of their maintenance is that coherence, adherence to common methodologies, standards and ontologies, and availability can be ensured. While a solution maintaining all the data at the institutions generating that data is feasible and provides flexibility and scalability, it does not guarantee adherence to the same methodologies, common data models and standards, or long term maintenance, which makes the use of that data by large communities of users more difficult.

Due to the very strict requirements for a production-level system, with available documentation and user support, commercial deployment and long term maintenance, we have also concluded together with BIG that current ACGT prototype tools and services cannot be directly used for the neoBIG project, however some of them could be part of a further targeted solution.

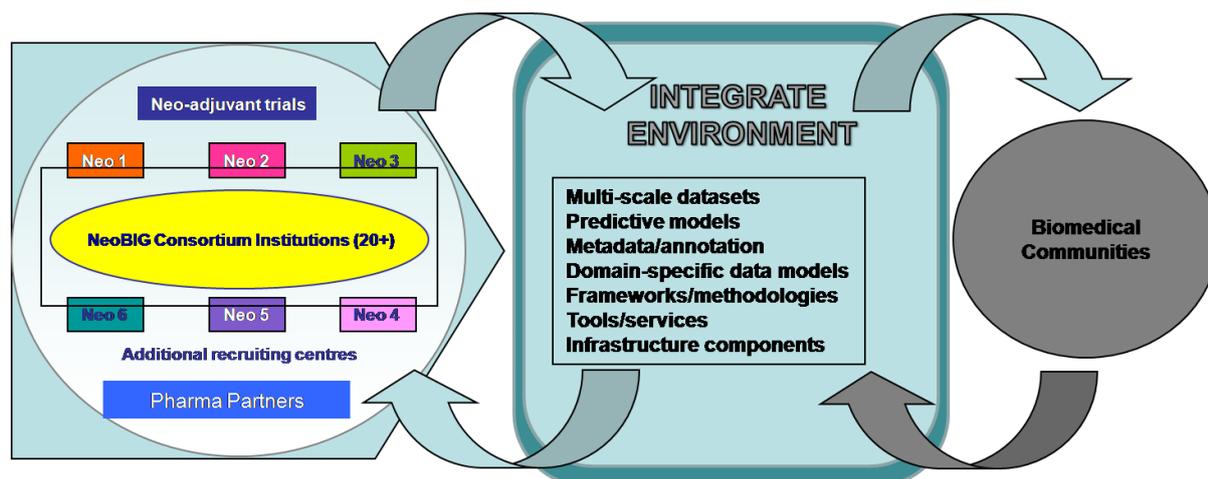


Figure 5 The INTEGRATE concept: Sharing and collaboration among clinical research and biomedical communities

In this context, we have defined INTEGRATE, a new collaborative project that aims to build solutions that support a large and multidisciplinary biomedical community ranging from basic, translational and clinical researchers to the pharmaceutical industry to collaborate, share data and knowledge, and build and share predictive models for response to therapies, with the end goal of improving patient outcome.

To address the needs identified during the neoBIG requirements analysis, the INTEGRATE project will develop flexible infrastructure components and tools for data and knowledge sharing and wide scale collaboration in biomedical research. Our infrastructure will bring together heterogeneous multi-scale biomedical data generated through standard and novel technologies within post-genomic clinical trials and seamlessly link to existing research and clinical infrastructures, such as clinical trials management systems, eCRFs, and hospital EHRs, and to relevant external biomedical infrastructures. We will also build repositories of data, annotated models, and metadata and provide tools to extract and manage content, add and update data

and models, and link to external sources for complex analyses. On top of this flexible infrastructure and using the available multi-level data, we will develop and validate models and simulators predicting therapy sensitivity for individual patients.

Next to bringing together data and knowledge, our solutions will join a wide multi-disciplinary community of biomedical and clinical researchers committed to work together, to establish common methodologies and clinical protocols, to collaboratively build predictive models, carry out research and select the most suitable integrative workflows. The infrastructure and tools developed by the INTEGRATE project will support BIG to promote in the clinical community new methodologies and define standards concerning the collection, processing, annotation and sharing of data in clinical research and improve the reproducibility of results of clinical trials.

INTEGRATE aims to build an environment providing to its users full support for collaboration and sharing of complex multi-level datasets and models, but also access to relevant external data, knowledge and services. At the same time, we aim to enable the biomedical research community to benefit of the comprehensive datasets preserved by the INTEGRATE environment, and of our predictive models and tools. We are aware that the value of our tools and infrastructure would be only limited without sufficient data. To this end, an important goal of the project will be to enable long term sustainability of the project solutions, with specific focus on long term maintenance of large datasets built with common methodologies and using standardized models and terminologies.

The above needs have been recognized world-wide with many high profile initiatives, such as caBIG¹ and Sage Bionetworks², having as a mission to bring together researchers and their data and knowledge, build tools and infrastructures enabling sharing and collaboration, support reuse of data, models and tools, and promote common standards and interoperability.

To maximize the impact of INTEGRATE, we have established collaboration channels to several large user groups such as the European Organization for Research and Treatment of Cancer (EORTC)³ and Sage Bionetworks. Sage is a new initiative in the US that aims to “revolutionize how researchers approach the complexity of human biological information and the treatment of disease”.

4.5 Other relevant initiatives – caBIG

An important relevant initiative is the cancer Biomedical Informatics Grid (caBIG)⁴, launched by the National Cancer Institute and maintained by the Center for Biomedical Informatics and Information Technology (CBIIT). The goals of caBIG are to connect scientists and practitioners through a shareable and interoperable infrastructure, develop standard rules and a common language to more easily share information, and build or adapt tools for collecting, analyzing, integrating, and disseminating information associated with cancer research and care.

¹ <http://cabig.cancer.gov/>

² <http://www.sagebase.org/>

³ <http://www.eortc.be/>

⁴ <https://cabig.nci.nih.gov/>

The basic infrastructure consists of two main parts. The *Cancer Common Ontologic Representation Environment* (caCORE)⁵ provides functionality to ensure interoperability. For instance, it includes a service for managing vocabularies (EVS) and a service for managing metadata (caDSR), along with a software development kit for easy access to these services. Typically, (object-oriented) information models are registered in caDSR and their meaning is linked to vocabularies stored in EVS. Furthermore, caBIG provides a technical platform and infrastructure (caGRID) based on GRID technology⁶. This allows for secure sharing of resources (for instance computational or data resources). These two parts form the basis for the more functional services, where typically a grid service is exposed expressed in terms of the caCORE.

caBIG provides various tools which further build on this basic infrastructure. Four *domain workspaces* were formed to focus development⁷:

- *Clinical Trial Management Systems*: Develops a comprehensive set of modular, interoperable and standards based tools designed to meet the diverse clinical trials management needs of the Cancer Center community.
- *Integrative Cancer Research Workspace*: Produces modular and interoperable tools and interfaces that provide for integration between biomedical informatics applications and data. This will ultimately enable translational and integrative research by providing for the integration of clinical and basic research data.
- *In Vivo Imaging Workspace*: Creates, optimizes and validates tools and methods to extract meaning from and share imaging data.
- *Tissue Banks and Pathology Tools Workspace*: Develops a set of tools to inventory, track, mine, and visualize biospecimens and related annotations from geographically dispersed repositories.

The INTEGRATE project will assess the relevant parts from caBIG from an adopt, adapt and interoperate perspective. Next to the functional requirements, the components should also fulfil various non-functional requirements (as will be specified during the use case elaboration), such as costs (e.g. various tools rely on Oracle Clinical), ease of deployment and maintenance, and performance.

Next, we highlight some relevant parts of caBIG. **caArray** provides an open array data management system which allows federation of multiple installations. **caTissue** is a biorepository tool for biospecimen inventory management, tracking, and annotation. **geWorkbench** provides an innovative, open-source software platform for genomic data integration, bringing together analysis and visualization tools for gene expression, sequences, protein structures, pathways, and other biomedical data. **Cancer Central Clinical Database** (C3D) is a clinical trials data management system. C3D collects clinical trial data using standard case report forms (CRFs) based on common data elements (CDEs). C3D utilizes security procedures to protect patient confidentiality and maintain an audit trail as required by FDA regulations. **The**

⁵https://cabig.nci.nih.gov/concepts/caCORE_overview?pid=primary.2006-07-07.4911641845&sid=caCORE_Overview&status=True

⁶ <http://www.globus.org/toolkit/>

⁷ <https://cabig.nci.nih.gov/workspaces/domain>

Clinical Connector⁸ provides a semantically integrated service layer via caGRID that allows C3D adopters to expose functions within C3D (Oracle Clinical). The exposed service layer uses a BRIDG based model and defines service operations that could be implemented by other CTMS systems. The first service allows external applications to load laboratory test result data into C3D study structures for registered patients. The second service allows external applications to 31nrol patients onto a C3D hosted study. Exposing additional C3D functionality is being planned.

For every component required in our project, we will assess whether there is a appropriate offering from caBIG. If the requirements do not match, a trade-off will be made between modifying an available component versus developing our own component. As various partners collaborate with non-European partners, emphasis will be put on standardization, harmonization and standards compliance of the solutions.

⁸ <https://cabig.nci.nih.gov/tools/C3DClinicalConnector>

References

- [1] Alberto Anguita et al (2010). D 7.9 – *Formal procedures and protocols for the semantic integration of clinical trials in ACGT*. ACGT
- [2] Antonioletti, M. E. (2005). The Design and Implementation of Grid Database Services in OGSA-DAI. *Concurrency and Computation: Practice and Experience, Volume 17, Issue 2-4* , 357-376.
- [3] Base. (n.d.). *BASE – BioArray Software Environment*. Retrieved from <http://base.thep.lu.se/>
- [4] Bizer, C. (n.d.). *D2RQ – Treating Non-RDF Databases as Virtual RDF Graphs*. Retrieved from D2RQ: <http://sites.wiwiwss.fu-berlin.de/suhl/bizer/d2rq/index.htm>
- [5] Bizer, C. (n.d.). *The D2RQ Platform v0.7 User Manual and Language Specification*. Retrieved April 2010, from D2RQ: <http://www4.wiwiwss.fu-berlin.de/bizer/D2RQ/spec/>
- [6] Bonsma, E . *D 5.7 -Guidelines and recommendations for integrating clinical data sources in the ACGT platform*. ACGT.
- [7] Brochhausen, M. (2007). *D7.2 – The ACGT Master Ontology*. ACGT.
- [8] Alan Cannon, et al (2004). *Ontology-Driven Automated Generation of Data Entry Interfaces to Databases*. Lecture Notes in Computer Science, 2004, Volume 3112/2004, 150-164, DOI: 10.1007/978-3-540-27811-5_15
- [9] Cocos, C. (2008). *D7.7 – Design Principles of the ACGT MO*. ACGT.
- [10] Phuong Dinh (2009). Presentation: *The NeoBIG program*. Pre-IMPAKT Training Course , May 6th 2009, Brussels.
Retrieved from <http://www.esmo.org/fileadmin/media/presentations/1324/opening/Dinh.ppt>
- [11] Jules Bordet Institute. Topoisomerase II Alpha Gene Amplification and Protein Overexpression Predicting Efficacy of Epirubicin (TOP). Retrieved from <http://clinicaltrials.gov/ct2/show/NCT00162812>; last visited: 7-18-2010.
- [12] International Society of Paediatric Oncology. *Nephroblastoma (Wilms Tumour) – Clinical Trial and Study SIOP 2001*. Final version January 2002, ammended 2004 and 2007, EUDRACT No.: 2007-004591-39.
- [13] Tamás Hauer, et al (2010). *An Architecture for Semantic Navigation and Reasoning with Patient Data - Experiences of the Health-e-Child Project*. Lecture Notes in Computer Science, 2010, Volume 5318/2010, 737-750, DOI: 10.1007/978-3-540-88564-1_47
- [14] Hildebrand, M (2006). */facet: A Browser for Heterogeneous Semantic Web Repositories*. *International Semantic Web Conference ((ISWC2006)*, (pp. 272-285). Athens.

- [15] Akrivi Katifori, et al (2007). *Ontology visualization methods—a survey*. ACM Computing Surveys (CSUR) archive. Volume 39 , Issue 4 (2007)
- [16] ObTiMA, Ontology based trial management application. Retrieved from <http://obtima.org/>
- [17] E. Patrick Shironoshita, et al (2008). *Semantic Representation and Querying of caBIG Data Services*. Lecture Notes in Computer Science, 2008, Volume 5109/2008, 108-115, DOI: 10.1007/978-3-540-69828-9_10
- [18] W3C. (n.d.). *Resource Description Framework (RDF)*. Retrieved from The World Wide Web Consortium : <http://www.w3.org/RDF/>
- [19] W3C. (n.d.). *SPARQL Query Language for RDF*. Retrieved from The World Wide Web Consortium : <http://www.w3.org/TR/rdf-sparql-query/>
- [20] Patrick Ziegler , Klaus R. Dittrich (2004). *Three Decades of Data Integration - All Problems Solved?* 18th IFIP World Computer Congress (WCC 2004), Volume 12, Building the Information Society
- [21] Zillner, S. Hauer, T. Rogulin, D. Tsymbal, A. Huber, M. Solomonides, T (2008). *Semantic Visualization of Patient Information*. Computer-Based Medical Systems, 2008. CBMS '08.

Appendix 1 - Abbreviations and acronyms

This glossary lists various acronyms that are used throughout the deliverable. It does, however, not include all acronyms that are used. Acronyms that are only introduced and used in a particular section, and not referred to subsequently, are typically excluded.

<i>API</i>	Application Programming Interface. The public interface provided by libraries and services.
<i>BASE</i>	BioArray Software Environment. The database for storing microarray data that is used within ACGT.
<i>CAT</i>	Custodix Anonymisation Tool. The tool used within ACGT to anonymise and pseudonymise clinical-trial data.
<i>CTMS</i>	Clinical Trial Management System.
<i>CRF</i>	Case Report Form.
<i>CSV</i>	Comma Separated Values. A simple textfile format for structured, tabular data.
<i>D2RQ</i>	A platform for accessing non-RDF, relational databases as virtual, read-only RDF graphs
<i>DICOM</i>	Digital Imaging and Communications in Medicine. A standard for exchanging medical data.
<i>DMS</i>	Data Management System. The grid-based file storage system that is used in ACGT.
<i>GAS</i>	Grid Authentication Server. The authentication server that is used within ACGT
<i>JDBC</i>	Java Database Connection. A Java API for database access
<i>LIMS</i>	Laboratory Information Management System.
<i>OGSA</i>	Open Grid Services Architecture
<i>OGSA-DAI</i>	OGSA standard for Data Access and Integration. A middleware product that supports the exposure of data sources onto the grid.
<i>PACS</i>	Picture Archiving and Communication System. Generic term for medical imaging databases.
<i>RDBMS</i>	Relational Database Management System.
<i>RDF</i>	Resource Description Framework. A language for representing

	information about resources.
<i>SCU</i>	Service Class User. DICOM term for the client of a service
<i>SCP</i>	Service Class Provider. DICOM term for the provider of a service.
<i>SPARQL</i>	A query language for RDF.
<i>SQL</i>	Standard Query Language. The most popular query language for relational databases.
<i>SVN</i>	Subversion, the version control system used within ACGT.
<i>URI</i>	Uniform Resource Identifier. A string of characters that identifies or names an object on the Internet. It is a generalisation of URL.
<i>URL</i>	Uniform Resource Locator. A type of URI that specifies where a resource is available, and the mechanism for retrieving it.
<i>XML</i>	Extensible Markup Language. The format that is used by web services to exchange data.