Formal procedures and protocols for the semantic integration of clinical trials in ACGT

Project Number: FP6-2005-IST-026996

Deliverable id: D7.9

Deliverable name: Formal procedures and protocols for the semantic integration of clinical trials in ACGT

Submission Date: July 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 7.9
Document name:	Formal procedures and protocols for the semantic integration of clinical trials in ACGT
Document type (PU, INT, RE)	PU
Version:	1.0
Submission date:	07/29/2010
Editor: Organisation: Email:	Alberto Anguita UPM aanguita@infomed.dia.fi.upm.es

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

ABSTRACT: This deliverable aims at describing the procedures and protocols that need to be applied when a new clinical trial is added to the ACGT infrastructure. The final goal of the described steps is to offer clinicians the possibility to run and control a trial exclusively within the ACGT platform and, eventually, be able to combine that trial's data with data from other trials. The presented procedures and protocols are based on the use of different tools and resources interacting with each other in order to make the mentioned features available to the user. An overview of those tools and resources, along with the involved procedures and protocols, is given in this document.

KEYWORD LIST: Clinical trials, Semantic Mediation, Ontologies, Wrapper

MODIFICATION CONTROL			
Version	Date	Status	Editor
0.1	07/20/2010	Draft	Alberto Anguita
1.0	07/29/2010	Final	Alberto Anguita

List of Contributors

- Alberto Anguita, UPM
- Fatima Schera, IBMT-FHG
- Holger Stenzhorn, USAAR
- Martin Dörr
- Mathias Brochhausen USAAR
- Ulf Schwarz USAAR

Contents

1 Executive Summary

This deliverable aims at describing the procedures and protocols that need to be applied when a new clinical trial is added to the ACGT infrastructure. A great effort inside the ACGT project was targeted at the construction of the semantic mediation layer and its related tools. With it, clinicians would be able to semantically integrate data from different trials, obtaining an important advance for their research. The ACGT platform includes several tools and modules that interact with each other to achieve such integration. One of them is the ACGT Master Ontology, covering the domain of clinical trials on cancer. This ontology provides the necessary semantic framework for the rest of components to achieve the semantic integration of different trials. Other tools focus on end-user interaction. For example, the ObTiMA system is the access point for end users to design and conduct new trials in the platform. The Semantic Mediator is the module in charge of integrating heterogeneous multi-level data. Along with the mentioned tools and modules, formal protocols and procedures have been designed to facilitate the actual semantic integration on new trials in the platform. An overview of the mentioned tools, and their related procedures and protocols, is given in this document.

2 Introduction

2.1 Purpose and structure of this document

This document describes the processes and protocols involved in the introduction of a new clinical trial in the ACGT technological infrastructure. In this section an introduction to post-genomic clinical trials on cancer is given. Section 3 describes the components and resources involved in the mentioned processes. Section 4 focuses on explaining the processes and protocols themselves. Finally, section 5 provides the conclusions of the suitability and viability of the developed processes and protocols.

2.2 Introduction

Modern clinical trials, and especially in the case of cancer research, rely on the integration of multilevel data to achieve new knowledge of disease behavior and therapy selection. Researchers focus on identifying genetic signatures that help predicting the goodness of a treatment for patients with similar genetic signature. For this purpose, both clinical and genomic data must be integrated and analyzed using complex data mining techniques. This approach has been applied successfully in a series of studies: genetic signatures for brain tumors [Petrik et al. 2006], genetic signatures for neuroendocrine cancer outcomes [Ippolito et al. 2005], gene expressions for Wilms tumors [Li et al. 2005], sets of genes associated to Nephroblastoma tumors [Zirn et al. 2006], gene signature for breast cancer metastases prediction [Wang et al. 2005], etc. The advantages of this type of approach are manifold: it enables the possibility of designing patient-specific therapies (also known as personalized medicine), helps avoiding unnecessary therapies and tests on patients which might do more damage than benefit (this is of special importance in the case on cancer diseases, where chemotherapy treatments can cause important effects on the patients health), help reducing costs, etc.

The ACGT technological infrastructure has been developed with the aim of facilitating the design and conduction of post-genomic clinical trials on cancer in an electronic manner. The advantages of this infrastructure over traditional procedures are manifold: computer based management of patient's data, automatization of time consuming tasks or analysis of CRFs data with advanced KDD tools among others. A feature of special relevance is the semantic integration of different clinical trials. With it, clinicians and researchers can access and analyze data from different clinical trials in a homogeneous manner—with the advantages that this implies, as described above. To support semantic integration, the ACGT Master Ontology, covering the domain of clinical trials in cancer, has been developed. This ontology provides the necessary semantic framework for an array of tools to work with disparate databases and offer the user the

desired integrated access. In this regard, several procedures and protocols have been designed to enable clinical trial integration in our platform. This document is devoted to the description of those procedures and the software components and resources that surround them.

3 The ACGT infrastructure for Clinical Trials

Several components of the ACGT infrastructure cooperate and collaborate to allow clinical trial chairmen to design and conduct new clinical trials in our technological platform. First, to provide the necessary semantic framework for clinical trials to be integrated between them, the ACGT Master Ontology, covering the domain of post-genomic clinical trials on cancer, has been developed. Next, to offer an intuitive platform for both designing and conducting clinical trials on cancer, the Ontology based trial management application (ObTiMA) and the Ontology Submission system have been created. Finally, to enable the semantic integration of different clinical trials introduced in the platform, the ACGT Semantic Mediation layer has been developed. These components interact between them to enable the desired features of cross-trial data analysis. They are described in the subsections below.

3.1 The ACGT Master Ontology

Technical Details

The ACGT Master Ontology (ACGT MO) is implemented in OWL-DL, the description-logics based subtype of the Web Ontology Language (OWL) [OWL] and can be freely downloaded from http://www.ifomis.org/acgt.

The initial development version of the ACGT MO was published in June 2007 and it has been further expanded since that time in order to integrate and respond to the needs of users, both clinical and technical. The developers are now working toward version 1.0. At the moment the ontology contains 1667 classes, 288 object properties, 15 data properties and 61 individuals. The ontology has been freely available since it was first published on the Internet in 2007, and comments and criticism of domain and ontology experts have been and are still invited.

There is currently an effort to reduce the number of object properties by around 60%. The reasons for this effort are both practical and principled. Practically speaking, it has become clear that 288 object properties are too many for most end-users to keep track of and utilize efficiently. On the other hand, from the standpoint of the ontology itself there are a number of redundant object properties, for instance undergoes_Process and undergoes_MedicalProcess.

Scope

The ACGT MO developers set out to comprehensively represent the domain of cancer research and management, with special emphasis on mammary carcinoma ("breast cancer"), Wilms' tumor (nephroblastoma) and rhabdoid tumor. The development of the MO was guided and reviewed by researchers from two pre-existing clinical trials, namely a breast cancer related trial on Topoisomerase II Alpha Gene Amplification and Protein

Overexpression Predicting Efficacy of Epirubicin (TOP) [TOP] and "Nephroblastoma (Wilms' Tumour) - Clinical Trial and Study SIOP 2001" by the International Society of Paediatric Oncology [SIOP]. In order to achieve the aim of supporting unified data annotation for these trials, the developers had to shape the MO as a cross-section of a multitude of sub-domains, all of which are vitally important to clinical cancer management and research. In effect, the outcome of this effort is best seen, not as a comprehensive domain ontology, but rather as an application ontology tailored to the needs of the ACGT users. A domain ontology is an ontology that has a clear-cut and distinguishable subject matter, one unified by the kinds of objects that it contains, by the dominance of a particular set of concepts and distinctions pertinent to these objects, and often by certain characteristic methods of inquiry as well. Paradigm examples of domain ontologies include representations of basic scientific subject matters, such as anatomy, cytology, the different areas of genetics, etc. The ACGT MO, by contrast, tackles a mixed bag of aspects arising from clinical cancer management, cancer research and clinical trial management. As a result of this, a single clearly delineated domain to which the ACGT MO applies cannot be easily identified. The MO, for instance, must represent administrative issues, as well as therapy- and laboratory-related facets of cancer in clinical reality. In designing it to do this we have been cautious to avoid the problem of use-mention mistakes that often occur in medical information systems. The use-mention distinction is violated when discourse that is intended to be about an object or kind of thing is phrased in such a way that it refers to the linguistic term for that thing rather than the thing itself. Consider the following two sentences:

1) Neoplasm is synonymous with tumor.

2) A neoplasm can be both, malign or benign.

The first statement is not a statement about neoplasms at all but rather a statement about the term “Neoplasm”, whereas the second is really a statement about actual things, namely neoplasms. Correctly formulated, 1) should be written as follows “Neoplasm” is synonymous with “tumor”. This example might seem relatively obvious, but in complex medical information systems statements about terms are quite often confused with or substituted for statements about the things in reality that the terms are intended to refer to. If an information system does not contain a sharp distinction between sentences of type one and type two, then consider what would happen if the system containing the above two sentences also contained the information: Neoplasm is a word. This would permit inference to the conclusion that there is some word that is either malign or benign, which is either false or, if true, not true in the same sense in which a neoplasm is malign or benign. So, a single use-mention confusion introduces either falsity or ambiguity into the information system, while many such confusions could truly compromise the overall quality of the data the system contains.

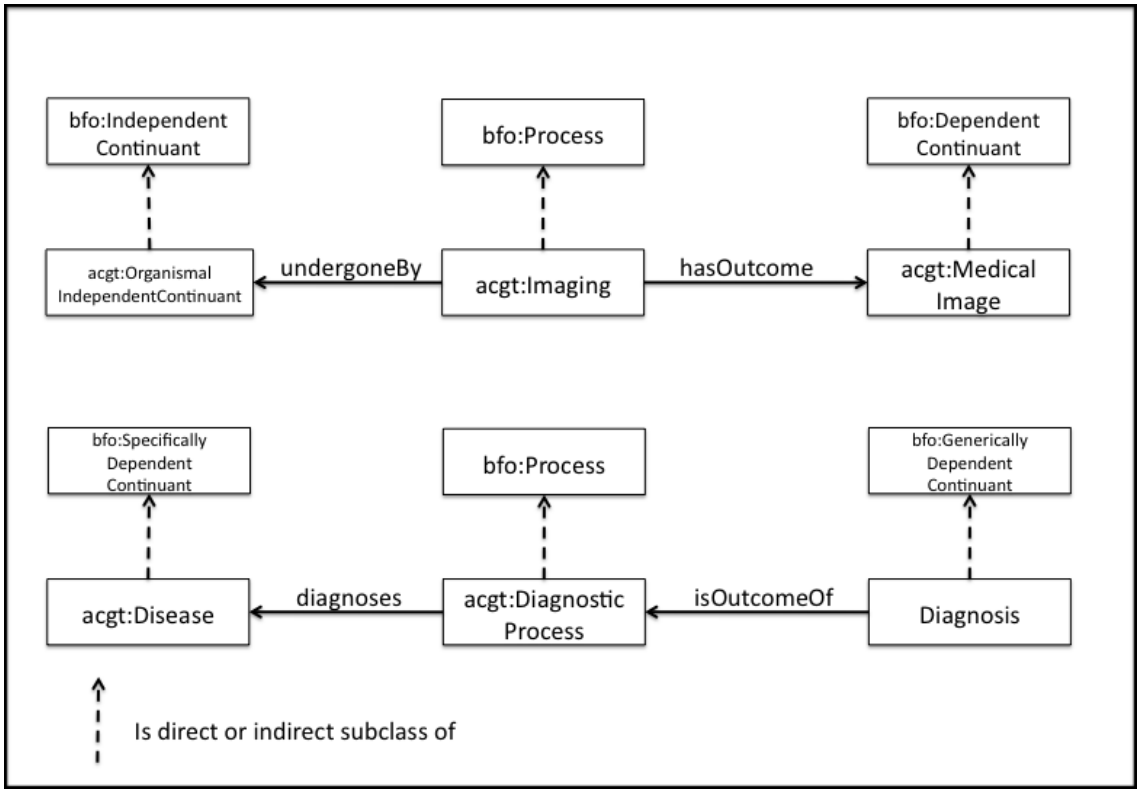

Fig. 1: Relations between specific information objects (Medical Image, Diagnosis) and processes, independent and other dependent continuants.

Thus, for the development of the ACGT MO it was crucial to avoid this kind of mistake, especially since we needed to represent both the clinical reality and the various kinds of documentation of clinical reality in the domain of our research. In order to guarantee this, our ontology includes a class called acgt:InformationObject, which includes items such as reports about entities, identifiers of entities and so on. ACGT is an extension of an upper ontology, namely Basic Formal Ontology (BFO) and we choose to make acgt:InformationObject a subclass of bfo:GenericallyDependentContinuant. A bfo:GenericallyDependentContinuant is defined as a continuant [snap:Continuant] that is dependent on some other independent continuant [snap:IndependentContinuant] bearer such that every instance of a generically dependent continuant D requires some instance of an independent continuant C, but which particular instance of C serves as the bearer of D can change from time to time [BFO]. For example, Leo Tolstoy's novel War and Peace (generically dependent continuant D) requires instantiation in some paper or electronic bearer (e.g. a book or a pdf file) C, but it is not particularly important for the existence of the novel as such which particular bearer instantiates it.

Examples of representations of detailed, real world clinical trial data are given in next subsection, where the Ontology-based Trial Management Application is described.

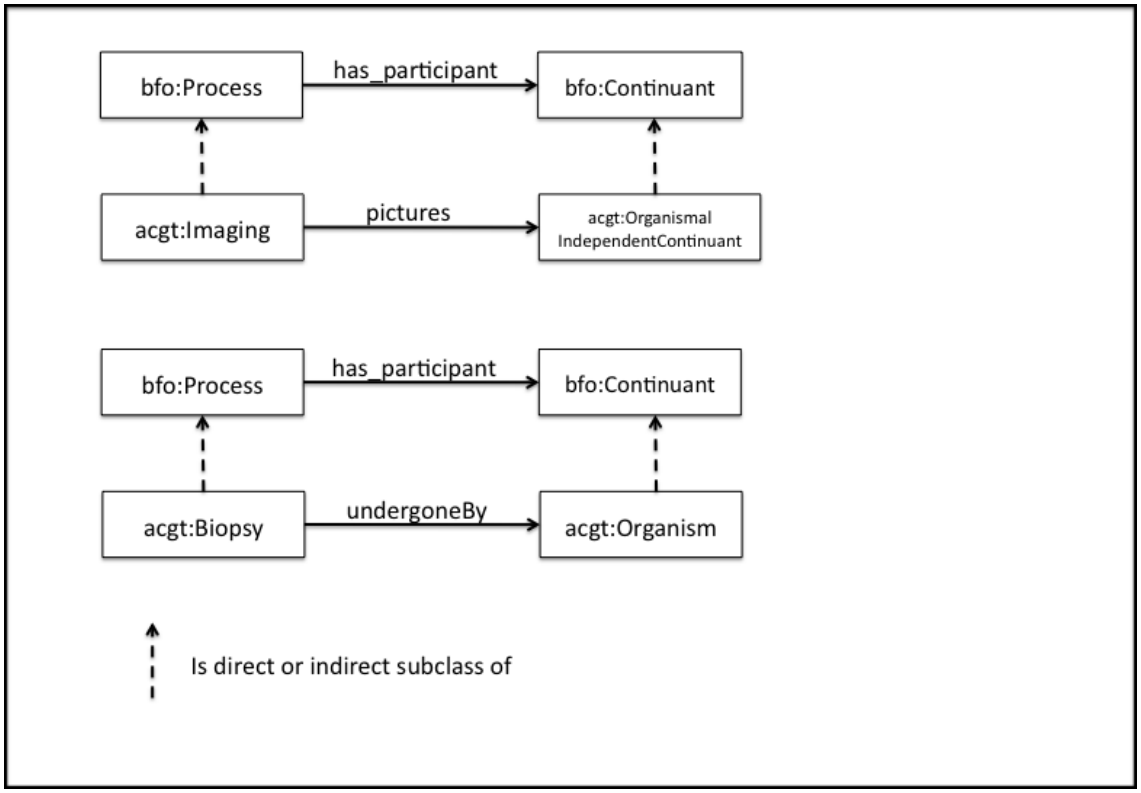

Fig. 2: Relations between ACGT-specific classes and their super classes from BFO.

Figure 1 shows a number of examples linking objects and processes from clinical reality to documentation items that are the results of these, as well as the subclass relation that each of these entities (the objects, processes and documentation items) stand in to various BFO classes. Figure 2 shows ACGT-specific relations as sub-relations of relations imported to the ACGT MO from an external source.

All these prerequisites make the ACGT MO an application ontology, one unified primarily by the goals or ends that it is designed to achieve or facilitate.

In what follows, we will show how the practical constraints introduced by real-world software development needs have interacted in innovative ways with the design principles that we hold to be necessary for high quality ontology development.

Aim

The ACGT MO is an application ontology and its main role, in the context of the translational medicine research framework within which it is developed and applied, is to support data integration across the borders of countries and disciplines, languages and professional terminologies; as well as integration of newly gathered data with existing data.

As a result, the ACGT MO is heavily used in the context of the ACGT Semantic Mediation Process. In specific, the two key systems exploiting

the MO are the ACGT Semantic Mediator and the Ontology-based Trial Management Application (ObTiMA).

The current version of ObTiMA aims to support clinical trial set up, design and management. In this context, the MO is utilized as a global schema for data annotation. We foresee that Version 2 of ObTiMA will include decision support with respect to many critical issues for clinical trial setup and management. Such functional requirements are, nevertheless, out of scope for the ACGT project and the development of this functionality will go hand in hand with a process of ontology development towards the needs of such services. As a conclusion, the ACGT MO does not aim to provide a comprehensive coverage of the complete domain neither in terms of class coverage nor in terms of class definition. Thus the development of new functionality and the expansion of the ontology itself are processes that will occur gradually and in tandem.

3.2 The ObTiMA System

Clinical Trial Management Systems (CTMS) promise to help researchers in hospitals and biotechnology/pharmaceutical companies to better manage the tremendous amounts of data involved when conducting clinical trials. Their goal is to simplify and streamline the various aspects of clinical trials, such as planning, preparation, performance, and reporting, by providing functionalities, like automatic deadline tracking for legal or regulatory approval, progress report issuing, keeping participant information up-to-date, or import/export data from/into other clinical information systems. For example, it is still a common yet tedious and error-prone practice to collect data at each trial site on paper-based Case Report Forms (CRF) and then to enter them manually into the trial database at the trial center. CTMSs are supposed to avoid this by providing user interfaces that blend into clinical work settings and shield users from underlying data and system complexity.

But as standardized, commercial CTMSs are not yet widely deployed, trial databases and their entry interfaces are often developed in-house specifically for a given trial and therefore not readily reusable in other trials. This issue causes an additional reimplementation burden and makes it difficult to compare or integrate data between different trials. But even if CTMSs are used, the following issue remains unresolved: Those systems allow a user to freely define the CRF items and structures without the need of any informatics skills. But although this is very desirable, it can create the same interoperability problems. If a database is derived from the trial-specific CRF definitions, the database in turn is again also trial-specific and data reuse in further research stays problematic. Thus, our work focuses on solving this interoperability issue through an approach based on ontology and semantic (data) mediation

3.3 The Ontology Submission tool

A major need of the ACGT community was to create a workflow and communication system that would gather all the change requests regarding the content of the ACGT MO, feed them to the ontology experts in a manageable way, keep the version history of the ACGT MO, and automate the communication back to the interested parties of any changes taken place. These functional requirements imply that the required information system should have the ability to reclassify content or to rewrite queries involving any authorized new expression that has replaced an old, an obsolete or a previously-used but currently rejected user-provided term. To that end the ACGT Submission System was created. The system is a reactive communication system allowing end users to criticize and/or submit their own opinion on the existing ACGT MO to its maintenance team.

The Submission System does not replace ontology development systems such as "Protégé". Rather, its role is to gather requests for changes, assist the ontology expert by providing access to those requests and by providing a point of reference for the changes in the ontology, and to maintain previous ontology versions on a per-class basis, including the history of related requests. The reason for this is simply that previous classes, versions of or changes to the ontology may well be of relevance in making future decisions about what to include or whether or not to make a change. The ACGT Submission System interfaces with an ontology development system, here Protégé, to implement changes in a particular version of the ACGT MO and to control the formal consistency of all classes in that version. It (semi automatically) traces and registers the changes made and relates them to previous versions of the ontology, including changes to individual classes and requests for such changes. The relatively loose coupling with Protégé has the advantage of rendering the ACGT Submission System highly generic and potentially useable with other ontology development systems in the future (Protégé, even though quite popular, is not yet stable enough to encourage a tighter coupling). The system manages the workflow of processing requests, the details of decision-making, and the necessary communications in order to minimize reliance on manual checking and carrying out of these things by human beings. It is inspired by the workflow patterns of well-known international thesaurus development teams such as the Getty Research Institution and English Heritage.

The Submission system can be accessed by authorized users independently through the Web or from within the ObTiMA System. Thus, ObTiMA users can add change requests to the ACGT MO directly from ObTiMA during the process of CRF document definition.

The ACGT Submission system distinguishes three user roles:

a) The Contributor. A contributor to the system is a person who wishes to comment or suggest changes to the ontology, requesting

additions/deletions or modifications of the existing ontology contents.

b) The Domain Expert. The Domain expert contributes to the system by reviewing the submissions of the Contributors that concern his field of expertise, and informs the Ontology Experts of the necessary changes to the ontology.

c) The Ontology Expert. The Ontology Expert is trained in logic and formal ontologies and in general possesses only minimal domain knowledge. (S)he is responsible for the maintenance of the ontology. (S)he receives all the change requests (submissions), answers them or forwards them to a Domain expert. This communication is automated to the highest degree possible.

The ontology experts can browse through submissions, review the submissions, discuss them with contributors and domain experts, and decide whether they agree or disagree with the proposed changes, leading to either their implementation or their rejection. Any rejection of a proposed change will be accompanied by a declaration of how the correct meaning of a proposed class is to be expressed by the MO (a migration path). In assistance, the system provides the ontology expert with adequate information services about all related class versions and submissions. The system provides automatic feedback in the form of notifications to the Contributors on the status of their submissions, and on the status of the ontology. The system manages the publication of sets of changes to the ontology on a release-by-release basis. A new release can be incorporated into the already running ACGT Information systems along with migration information.

3.4 The ACGT Semantic Mediation layer

Designing a clinical trial (CT) with the ObTiMA system and inputting data in its CRFs is just the first step to introduce it in the ACGT technological platform. For the CT to be semantically integrated with other CTs, it must be configured in the semantic mediation layer. This layer offers other tools and end users the possibility to query and retrieve data from the CRFs of the CT. The reason for developing a complete software layer for this task is performing the data access in terms of the ACGT MO—i.e. all queries to data in the CRFs will be done using classes and properties contained in the MO. This approach requires a big developing effort, but offers important advantages. First, data is mapped to a sound and solid semantic framework which ensures a correct data definition. Second, end users are not forced to learn the specifics of each CT datasource, as they will all be accessed in the same manner. Third, by sharing the MO as schema for all CTs, users can perform integrated queries across different CTs in a transparent manner.

The Semantic Mediation layer comprises several components that interact with each other to offer the described features. First, we have the Data

Access Layer, comprised of the database wrappers—software modules that translate any underlying database so it can be accessed using SPARQL [SPARQL] language. Above these we have the Semantic Mediator core, which is in charge of receiving SPARQL queries expressed in terms of the MO and translating them into terms of the underlying database vocabulary, and for subsequently aggregating all separate results into an integrated result set. For the semantic mediator to be able to translate queries between terms of the MO and terms of the source DBs, there needs to be a map between the two schemas. The mapping data is stored in the database mappings, which are XML documents describing the relation of a database with the MO. The mapping format describes which elements can be found in the mapping files, and covers all possible cases of semantic heterogeneity between two schemas.

Another important tool comprising this layer is the ACGT Query Tool. This is a web-based graphical tool aimed at end users, which allows building queries avoiding the technical details of the SPARQL query language. The functioning of this tool is described in detail in section 4.4.

4 Clinical Trial integration process

This section contains descriptions of the processes and protocols involved in the introduction of a clinical trial in the ACGT semantic infrastructure. Figure 3 depicts this process, with the red areas indicating the most complex/effort demanding tasks.

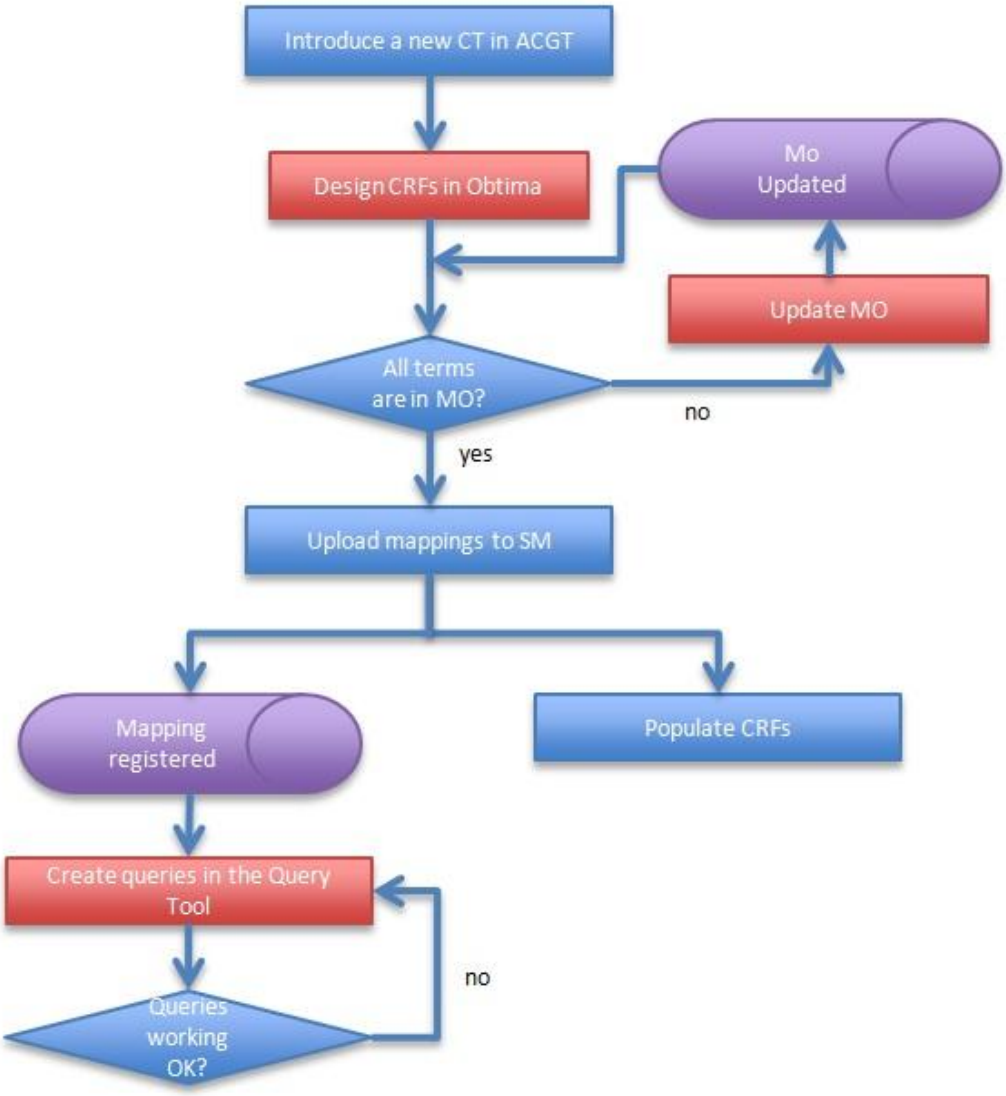

Fig. 3: Steps involved in the semantic integration of a clinical trial in the ACGT infrastructure. Red elements indicate the most effort-consuming tasks.

Different actors are involved in this process, beginning with the trial chairman, who must undertake the task of designing the trial in the ObTiMA system. In case the chairman finds any missing terms in the MO, the designated ontology expert will have to analyze and resolve the issue. The different procedures and protocols involved in this process are described in more detail in the next subsections.

4.1 Design of CRFs in ObTiMA

Trial Builder

The Trial Builder represents one of ObTiMA's two main components (Fig. 4) and enables the user to specify the various aspects of a clinical trial. The trial metadata can be defined in a master protocol based on templates for describing the trial goals and its administrative data, like start or end date. Treatment plans can be graphically designed to guide clinicians through the treatment of individual patients, and particular treatment events—such as chemotherapy or surgery—can be defined with all necessary information. The particular order of treatments for individual patients can be defined by placing them on a timeline. Also, treatment stratifications and randomizations to be applied for a patient can be described. For each stage on the treatment plan a CRF can be assigned to collect the data documenting the treatment and treatment outcomes.

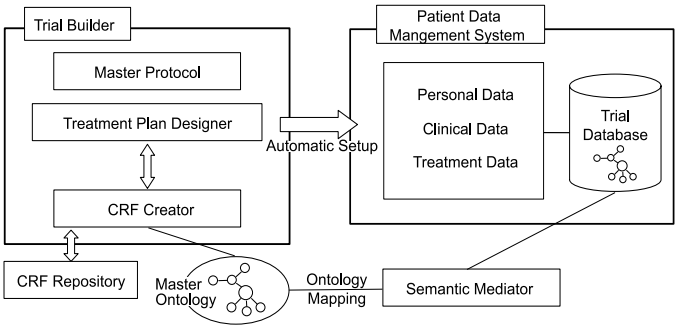

Fig. 4: ObTiMA System Components.

Ontology-based CRF Creation

The creation of CRFs marks the core functionality of the Trial Builder. In a graphical user interface, the user can define the content, layout, and navigation of the CRFs which are used to capture all patient data during a clinical trial, like the patient's history, medical findings, diagnostic data, or genomics data.

It is important that all information can be defined here which are necessary for the data integration, i.e., each CRF item is described based on ontology concepts together with metadata, like data type and measurement unit, to set-up the trial database. However, the internal CRF (data) representation is not the focus of clinicians but their "user interface" (layout) and their adaption and integration into the specific workflow of the planned trial: clinicians are not to be bothered with the underlying aspects of the trial database or the ontological metadata. Thus all these aspects are made transparent to the user through a graphical

user interface which hides the actual complexity yet gathers all required information for automatically creating the trial database. This interface is derived also automatically from the content and structure of the Master Ontology but represents a simplified ontology view, adapted to the task of creating items (Fig. 5). It comprises the following sections:

In the *Ontology View* (1), the user selects concepts from the ontology to create a CRF item. Here, the interface tries to overcome the gap between clinical practice and the actual logical representation of ontology concepts: Although the ontology provides natural language descriptions for its concepts/relationships (in addition to the logical definitions), those often do not fully mirror the needs of practical or clinical perception of reality. In order to meet this need, we do not present the full Master Ontology here but rather a simplified clinical view which contains a trial-independent basic classification of CRF contents from a clinician's point of view.

It is by intention that the clinical view is far less detailed as the actual Master Ontology and since this allows the possibility to provide a much easier entry point for the user. The interface of the clinical view is implemented as a tree always that starts at node of the concept "Patient" as focus of any clinical study (and hence CRF) and only presents those concepts that are directly reachable from this concept, like "Weight" or "Tumor" (indicating a patient's tumor). Only when a concept is selected then also the concepts directly reachable from this one are shown, such as "Laterality" in the case "Tumor" was initially chosen (indicating the laterality of the patient's tumor).

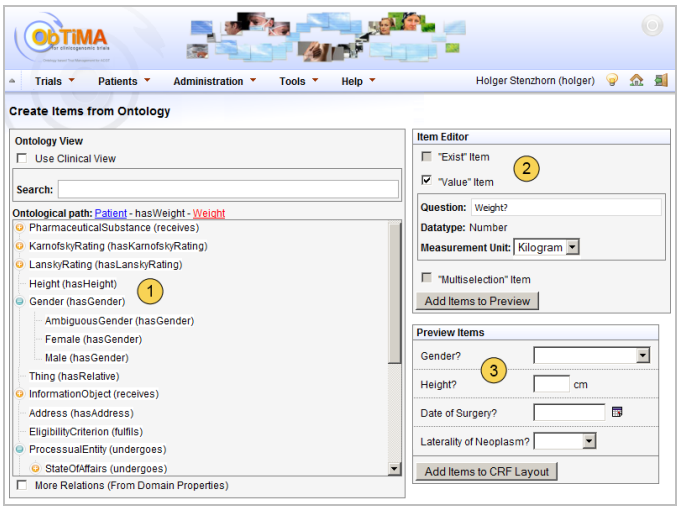

Fig. 5: Ontology Viewer while Creating CRF Items.

When a concept is chosen in (1) then a corresponding item is automatically created and shown in the *Item Editor* (2) together with its attributes determined automatically based on the chosen concept, such as label, data type, or answer possibilities, and which can be manually adopted. For example, the concept "Weight" has a numerical data type

and a list of suitable measurement units attached. So, when the CRF with this item is used in a clinical trial then the measurement units are offered as selection possibilities (in a drop down menu). The specified value (entered into a text field) is automatically tested to be of numerical type and also to be non-negative (since a weight cannot be negative). Finally, *Preview Items* (3) presents all created items in the order in which they are intended to appear on the CRF. Single items can be reordered by simple drag and drop and subsequently transferred to the interface where the overall layout of the CRF is then defined in turn.

CRF Repository

Revisiting the reuse and interoperability issue discussed in the introduction, in many trials similar or equal data are collected, yet stored differently because of different data(base) definitions. Applying the Master Ontology already improves this situation through using standardized concepts when creating CRFs. Going a step further, the situation would be further improved by partial or complete reuse of existing CRF in case similar data is collected. This idea has been realized by creating a unified CRF Repository as crucial part of ObTiMA. This repository allows the storage and retrieval of entire ontology-based CRFs and single CRF items or components for reuse and adaption in subsequent trials: When setting-up a clinical trial, applicable CRFs can either be directly reused or new ones quickly created by "plugging together" existing CRF items and components. This in turn fosters the standardization of CRFs even more, since CRFs can now be compared not only on the level of single items (through their basis on ontological concepts) but also on the level of larger components or in their entirety.

Patient Data Management System (PDMS)

The PDMS supports clinicians when conducting a clinical trial and is automatically set-up based on the master protocol and CRFs defined in the Trial Builder. The PDMS guides the clinicians through the actual treatment of patients according to their individual treatment plans and provides a graphical user interface to fill in the CRFs relevant to the patient's current treatment situation. The interface is adjusted to everyday clinical needs: As with the Trial Builder, the complexity of the underlying ontology is hidden from the user, yet its logic-based concept definitions are used to provide direct validity checking when CRFs are filled in. The basic look of the data entry interface corresponds to section (3) on Figure 5 with each input element providing on-the-fly feedback about its current state based on the just mentioned checking, i.e., in case a negative value is specified for a weight then this error is immediately highlighted along with an explanation of the error.

Data Export

To integrate ObTiMA into real-world clinical settings, the system must be capable to interface with other existing CTMSs and be able to exchange data in a format they understand. To meet this requirement, ObTiMA allows to import and export trial metadata, CRF descriptions and patient data through an extended version of the CDISC Operational Data Model (ODM) format [Kuchinke et ak. 2006]. This platform-independent, quasi-standard for exchanging and archiving clinical trial data is supported by many current CTMSs. Observing CDISC's extension guidelines, we enriched this format by allowing the additional inclusion of (metadata) descriptions based on Master Ontology concepts. In the case other CTMSs want to import data generated by ObTiMA, they can chose to interpret the supplemental descriptions but if this is not feasible the resulting data is still conform the ODM format and can sensibly be used by those systems.

Administration, Security and Pseudonymization

To administer multicentric clinical trials, ObTiMA contains several advanced facilities for managing the multitude of institutions, researchers, and patients usually participating in such trials. An elaborated, fine-grained security architecture has been implemented to handle the rights and roles that can be attached to the system's users in order to guarantee that they can only perform the tasks which they are fully authorized for. It is also straightforward to dynamically react to changes within a running clinical trial, since new institutions and users can always be added or extra security roles and rights can be defined.

It is also indispensable that ObTiMA, as a system holding real patient data, securely stores all of the data—which could possibly identify some patient to non-authorized persons—in pseudonymized and encrypted form. To foster security even more, additional security features will be included in the final version. Personal data is physically separated from the actual clinical research data through the use of two distinct database servers: One server holds the database for storing the personal data of the patients, such as their names and addresses (which must never be shared, e.g., via the Semantic Mediator). The protection of this database strictly follows all current legal regulations for data protection in clinical environments. The other server hosts the database that contains the actual research data collected in a clinical trial (through the use of the CRFs). It is possible within the Trial Builder to mark certain CRF items as personal which results in this data being stored in the database for personal data and not in the one for research data.

4.2 Updating of MO terms

In this subsection the process following a new submission to the MO (Figure 6) is described in more detail:

When inserting a new change request (submission) into the System, the End User automatically receives a notification certifying the submission. Once this is done, the new submission is inserted into the submission pool of the System. These new submissions are sent via mail to the Ontology Expert (a team or an individual), in order to inform her about the new change requests, and the Ontology Expert can see the new submissions to the system by logging into the system.

In the sequence, the Ontology Expert reviews the new submission. The submission may be directly accepted, being seen as redundant, or the Ontology Expert may need domain expert advice. If it is accepted, the contributor receives a notification. It is redundant if it refers to something already covered by the MO. In such a case it is rejected along with an explanation. If more domain expertise is needed, the Ontology Expert sends the submission to the Domain Expert (a group or individual). The Domain Expert will be informed via mail about the submission. After the Domain Expert has checked the submission, he can either reformulate it and send it back to the Ontology Expert or introduce an Implementation Proposal for the request. Either way, the Domain expert sends the submission back and the Ontology Expert accepts, reject, or postpones the submission and sends an answer, i.e., the way it will be implemented or not implemented, to the Contributor. At release time, all contributors are once again notified that their accepted submissions have been released in an authorized version.

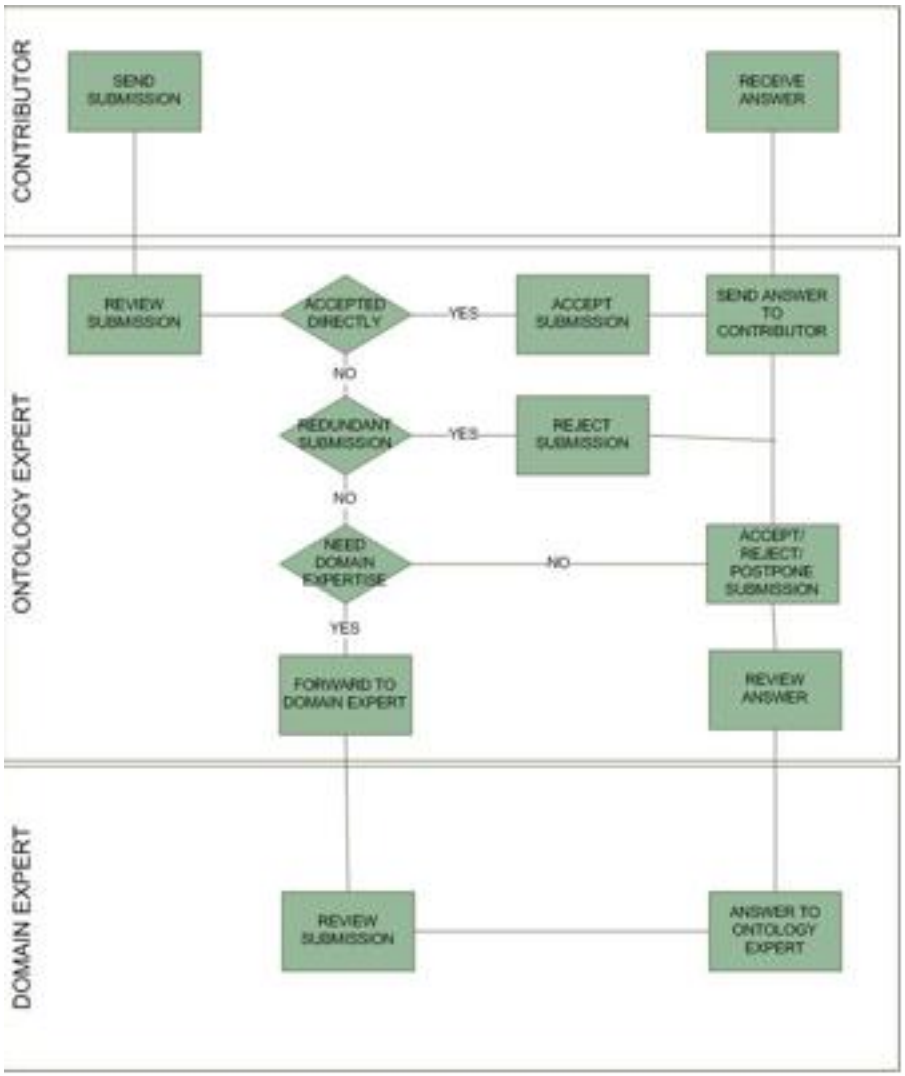

Fig. 6: The Submission Process.

Detailed documentation on how to use the Ontology Submission System can be found in deliverable D7.6, available at https://bscw.ercim.eu/bscw/bscw.cgi/99073.

4.3 The mapping process

The goal of the mapping process is the production of a "mapping file"—i.e. a set of correspondences between the MO and an underlying database schema. A correspondence is a pair of semantically equivalent elements in both schemas. In the ACGT approach, the queries are built in terms of the information contained in the mapping files.

The mapping process for legacy databases is a manual process that requires the involvement of a team of experts in different domains—namely, an MO authority, an expert in the database and an expert in mappings. In the case of CTs designed with the ObTiMA system, however, this process is completely automatic. An API for programmatically creating mapping files has been created so that ObTiMA can produce the mapping

at the same time that the CRF is designed. The generated mapping files are dynamically submitted them to the semantic mediator, so the newly created CT is immediately available in the data mediation layer.

4.4 Creation of queries in the Query Tool

As described in previous sections, the last step for enabling the use of a CT data resource in the ACGT infrastructure is to create the appropriate queries for it. The queries will be necessary for subsequent workflows and data analysis tools to be able to process the CT data.

The ACGT Query Tool allows creating queries in SPARQL language—the query language accepted by the Data Access Layer tools—in an easy an intuitive manner. It has been designed to allow researchers and clinicians lacking advanced technical background to seamlessly create queries that fit their needs. Through this graphical tool they will be able to explore the available elements to be queried and select which ones they want to retrieve, with the possibility of including advanced filters and restrictions.

The first step when using the Query Tool is to select one of the available repositories (each one corresponding to a data source integrated in the ACGT infrastructure). With the selected repository, a series of entries—i.e. fields of the data source that can be retrieved, like for example patient's birth date—will be listed. By selecting some of these entries we will compose a query. Note that entries from more than one repository can be incorporated to a query, allowing the homogeneous retrieval of data from disparate sources. After all fields to retrieve have been added to the query, we will be given the option to add additional restrictions to the query, or limit the number of results retrieved. Figure 7 depicts the construction of an example query using the Query Tool.

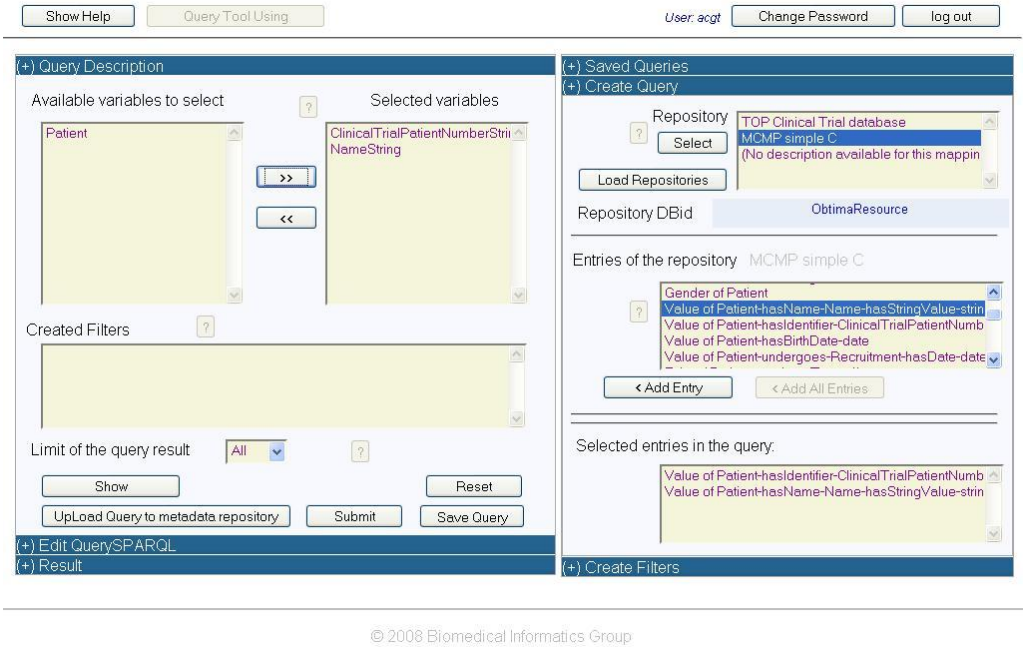

Fig 7: The Query Tool allows easily building SPARQL queries for the ACGT Data Access Layer.

After a query has been created, it can be tested by submitting it against the ACGT Semantic Mediation layer. The results of the query will be displayed in an additional screen, as shown in figure 8.

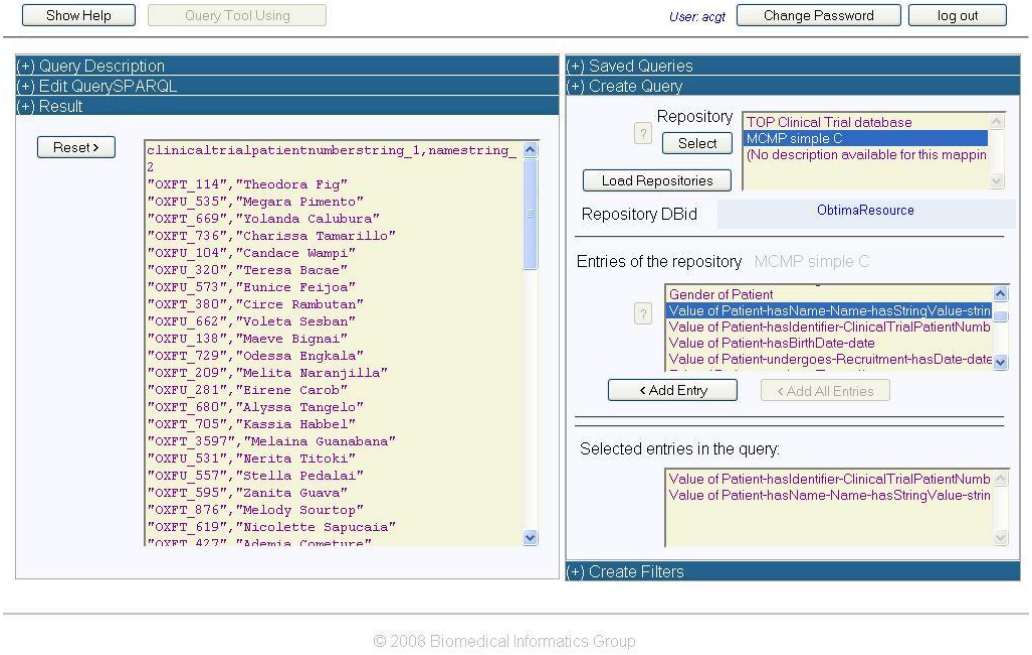

Fig 8: The Query Tool allows submitting created queries and displays the retrieved results in a separate window.

Finally, the queries can be submitted to the query repository. Queries stored in this repository can be included in complex workflows that combine data access operations with data processing and analysis.

5 Conclusions

During its four and a half years of duration, the ACGT project has focused on delivering an advanced infrastructure for conducting and managing clinical trials on cancer. One of the main features of this infrastructure was the semantic integration of clinical trials. The goal was to allow clinicians and researchers to access and analyze data from different clinical trials in a homogeneous manner—with the advantages that this implies, as described in section 2. To support semantic integration, the ACGT Master Ontology, covering the domain of clinical trials in cancer, has been developed. This ontology provides the necessary semantic framework for an array of tools to work with disparate databases and offer the user the desired integrated access. In order to use these tools, a series of formal procedures and protocols have been designed. These procedures are, namely: i) the CT definition in the ObTiMA trial builder system, ii) the submission of new terms to the ACGT Master Ontology—optional, only needed of missing terms are identified—, iii) the mapping of the generated CRF database to the Master Ontology and iv) the construction of integrated queries through the ACGT Query Tool. These tasks are automated to some extent, and in part require interacting with the end user. Only the process of generating mappings is completely automatic, and is performed in the background without the user noticing. Web-based tools with graphical interfaces have been developed to help and guide the user in the rest of the tasks. These tools have been tested with clinicians, providing very satisfactory results. The users were able to design new clinical trials with little effort—and much less when compared to a "traditional" process of design of a clinical trial.

The tests and experiments performed in demos with the described tools show that the goal of providing and infrastructure for creating and managing clinical trials, and in addition enabling their semantic integration, has been fully achieved. Integration of multicentric and multilevel data is crucial in present and future research in cancer-related clinical trials. Clinicians expect this approach to allow identifying genetic signatures that help selecting the most appropriate treatment for patients. Systems that perform transparent semantic integration of sources have been in development for over a decade. In contrast, less effort has been dedicated to facilitating their access to end users. The presented formal procedures and protocols—and their inclusion in the ACGT technological infrastructure—allow clinicians and researchers accessing these technologies and take full advantage of them.

6 Acknowledgements

Parts of this document are based on the paper published by Brochhausen et al. in Journal of Biomedical Informatics in May 2010, titled "The ACGT Master Ontology and its Applications – Towards an Ontology-Driven Cancer Research and Management System" (e-published ahead of print, DOI: 10.1016/j.jbi.2010.04.008).

7 Bibliography

[BFO] Basic Formal Ontology (BFO). Available from http://www.ifomis.org/bfo; last visited: 7-18-2010.

[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

[Kuchinke et ak. 2006] Kuchinke W, et al. Extended cooperation in clinical studies through exchange of CDISC metadata between different study software solutions. Meth Inf Med. 2006;45(4):441-6.

[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.

[OWL] OWL Web Ontology Language Semantics and Abstract Syntax. Available from http://www.w3.org/TR/owl-semantics/; last visited: 7-18-2010.

[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

[SIOP] 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.

[SPARQL] SPARQL Query Language for RDF. Available from http://www.w3.org/TR/rdf-sparql-query/; last visited: 7-18-2010.

[TOP] http://clinicaltrials.gov/ct2/show/NCT00162812; last visited: 7-18-2010.

[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

[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.