Applications of the ACGT Master Ontology on Cancer

Mathias Brochhausen[1], Gabriele Weiler[2], Luis Martín[3], Cristian Cocos[1], Holger Stenzhorn[4], Norbert Graf[5], Martin Dörr[6], Manolis Tsiknakis[6] and Barry Smith[1,7]

[1] IFOMIS, Saarland University P.O.15 11 50, 66041 Saarbrücken, Germany,
[2] Fraunhofer Institute for Biomedical Engineering, St. Ingbert, Germany
[3] Biomedical Informatics Group, Artificial Intelligence Laboratory, School of Computer Science, Universidad Politécnica de Madrid, Madrid, Spain
[4] Institute of Medical Biometry and Medical Informatics, University Medical Center, Freiburg, Germany
[5] Paediatric Haematology and Oncology, Saarland University Hospital, Homburg, Germany
[6] Foundation for Research and Technology-Hellas (FORTH), Institute of Computer Science, Heraklion, Greece
[7] Department of Philosophy and New York State Center of Excellence in Bioinformatics and Life Sciences, University at Buffalo, USA
Mathias Brochhausen, mathias.brochhausen@ifomis.uni-saarland.de

Abstract. In this paper we present applications of the ACGT Master Ontology (MO) which is a new terminology resource for a transnational network providing data exchange in oncology, emphasizing the integration of both clinical and molecular data. The development of a new ontology was necessary due to problems with existing biomedical ontologies in oncology. The ACGT MO is a test case for the application of best practices in ontology development. This paper provides an overview of the application of the ontology within the ACGT project thus far.

Keywords: biomedical ontology, clinical trials, mediation

1 Introduction

Over the last decade the amount of data on cancers and their treatment has exploded due to advances in research methods and technologies. Recent research results have changed our understanding of fundamental aspects of cancer development at the molecular level. Nevertheless, irrespective of the fact that huge amounts of multilevel datasets (from the molecular to the organ and individual levels) are becoming available to biomedical researchers, the lack of a common infrastructure has prevented clinical research institutions from being able to mine and analyze disparate data sources efficiently and effectively. As a result, very few cross-site studies and multi-centric clinical trials are performed, and in most cases it is not possible to seamlessly integrate multi-level data. Moreover, clinical researchers and molecular biologists often find it hard to take advantage of each other's expertise due to the absence of a cooperative environment which enables the sharing of data,

resources, or tools for comparing results and experiments, and of a uniform platform supporting the seamless integration and analysis of disease-related data at all levels [1]. This situation severely jeopardizes research progress and hinders the translation of research results into benefits to patients.

The Advancing Clinico-Genomic Trials on Cancer (ACGT) integrated project aims to address this obstacle by setting up a semantic grid infrastructure in support of multi-centric, post-genomic clinical trials [2]. This system is designed to enable the smooth and prompt transfer of laboratory findings to the clinical management and treatment of patients. Obviously, this goal can only be achieved if state-of-the-art semantic technologies are part of the IT environment. In order to meet this goal, the ACGT project needed an ontology to be utilized in the context of its selected Local-As-View (LAV) data integration strategy [3]. In such a strategy the ontology plays the role of a global schema to which all local schemata are mapped, so that all their mapped equivalents are subsumed by the global schema. This requires that the global schema (i.e. the ontology) be sufficiently generic as to cover not only terminology, but also the meaning of all local schema constructs. The ACGT project achieves the semantic integration of heterogeneous biomedical databases through a service oriented, ontology driven mediator architecture that makes use of the ACGT-MO [4, 5]. The new terminology resource which underlies this integration rests upon a thorough review and critical assessment of the state of the art in semantic representation of cancer research and management.

2 Pre-Existing Ontologies and Terminologies

Cancer has been a focus of interest in biomedical research for a very long time. As a result of this long history, a number of terminological resources exist that are of relevance to ACGT. In order to prevent redundancy, the project undertook a very detailed review. We will illustrate this selection process by focusing on two potential resources that did not meet our criteria of excellence, and hence were either not used in ACGT, or were used after considerable alteration. We will further mention two general biomedical resources selected for integration in the ACGT terminological network.

When considering the development of an ontology-based information-sharing system for the cancer domain of the sort used by ACGT, the National Cancer Institute Thesaurus (NCIT) is a terminology resource of obvious relevance [6]. Yet, there are a number of drawbacks preventing the use of the NCIT as semantic resource of the ACGT project, in part because its formal resources are too meager for our purposes, with only a fraction of NCIT terms being supplied with formal definitions of the sort required by its official description logic (DL) framework. The NCIT contains only one relation, namely the subtype relation (*is_a*), as contrasted with the plurality of formally defined relations included, for example, within the OBO Relation Ontology [7]. Further, the NCIT is marred by a number of problems in its internal structure and coverage [8], including problems in the treatment of *is_a*. For a quick illustration of the inadequate treatment of *is_a* in the reviewed version of NCIT. let us consider the NCIT class *Organism*, which includes among its subtypes *OtherOrganismGroupings*;

with this we have *OtherOrganismGroupings is_a Organism* [6]. Given the formal definition of the subtype relation this is clearly wrong; groupings of organisms are not themselves organisms.

Another resource that has the aura of indispensability in a domain dealing with gene array data is the Microarray and Gene Expression Data (MGED) ontology [9]. Yet, even this highly used resource shows considerable deficiencies, including informal *is_a* relations. The inconsistency becomes obvious when the textual definitions – which are an asset to MGED – are taken into account: According to the MGED ontology *Host* is a subclass of *EnvironmentalHistory*. It is obvious that this cannot be a formal *is_a* relation. Taking a close look reveals an astonishing incoherence here: The definition of *Host* is: "Organisms or organism parts used as a designed part of the culture (e.g., red blood cells, stromal cells)" [9]. The definition of *EnvironmentalHistory* reads as follows: "A description of the conditions the organism has been exposed to that are not one of the variables under study" [9]. The thesis that an organism or organism part is a description clearly involves a crude category mistake (the confusion of use and mention). For some portions of the ACGT domain, however well-built and well maintained ontologies with high usability could be identified and reused within ACGT. This, as a matter of fact, applies both to the Foundational Model of Anatomy (FMA) [10] and the Gene Ontology (GO) [11], since they both fulfill the requirements on coherence and theoretical rigor specified in [12].

Most of the current ontologies for life sciences start from terminology appearing in documentation systems as data and pertaining to the "subject matter" of the research carried out, such as concepts about the human body, diseases and microbiological processes. However, the data kept in the systems ACGT aims at supporting also pertain to the scientific processes of observation, measurement and experimentation together with all contextual factors. A model for integrating that data must include this aspect. The CIDOC CRM (ISO21127, [13]) is a core ontology originally developed for schema integration in the field of documenting the historical context and treatment of museum objects, including a generic model of scientific processes. Some concepts and relations of the latter were reused and refined for the ACGT MO.

Effectively, developing a new ontology was imperative, since no single ontology or set of ontologies had the respective coverage and logical consistency.

3 The ACGT Master Ontology

3.1 Technical Details of the ACGT MO

The intention of the ACGT MO [4] is to represent the domain of cancer research and management in a computationally tractable manner. As such, we regard it as a domain ontology. The initial version of the ACGT MO that was made public on the internet consists of 1300 classes. The ontology was built, and is being maintained, using the Protégé-OWL open-source ontology editor [14]. It is written in OWL-DL [15] and presented as an .owl file. The ACGT MO not only represents classes as linked via the basic taxonomical relation (*is_a*), but connects them via other semantic relations called "properties" in OWL terminology. The OBO Relation Ontology (RO) [7] has

been used as a basis in this regard, as RO has been specifically developed to account for relations in biomedical ontologies [16]. Some properties of scientific observation were taken from the CIDOC CRM [13].

3.2 Methodology

The ACGT MO has been developed in close collaboration with clinicians utilizing existing Clinical Report Forms (CRFs), which were used to gather documentation on the universals and classes in their respective target domains, and to understand the general semantics of form-based reporting of clinical observation. All versions of the ontology have been reviewed by clinical partners who have proposed changes and extensions according to needs. In this process the problem of handling an ontology with more than 1300 classes for clinical users became apparent. Providing tools to examine the ontology in user-friendly ways emerged as inevitable. Yet, to ensure comprehensiveness of the representation of relevant portions of reality it was found necessary to go beyond the CRFs and the documentation provided by the clinical project partners. The latter governed the development of the leaf nodes of the ACGT MO, but we had to identify classes for a middle layer of the representation in order to ensure that the ontology provided the necessary reasoning support. Therefore, standard literature and standard classification systems were used, e.g. [17, 18, 19]. In order to provide a consistent and sound representation, the ACGT MO employs the resources of an Upper Ontology, which does not represent domain specific knowledge, but consists of classes that are generic and abstract [20]. The ACGT MO is based on Basic Formal Ontology (BFO) [21], which has proven to be highly applicable to the biomedical domain [22], and is now providing the advantage of common guidelines for ontology building to a multiplicity of research groups and organizations,

It is a well-documented fact that well-built, coherent ontologies tend to be hard to understand for clinical users [23]. An *is_a* hierarchy based on BFO puts kinds of processes and kinds of objects on quite distant branches. The clinician should, nevertheless, have these associations readily available on the screen. We therefore proposed that the basis for these tools should be a viewing mechanism that should reflect the terms typically appearing together in particular clinical contexts, while the full ontology was running behind the scenes. The necessary associations may be found and activated by tracking the workflows commonly used in computer applications serving clinical practice. In the following we present several specific techniques and work styles that were employed in the development of the ACGT MO.

Lassila et al. [24] categorized ontologies according to the amount of information they contain. Their classification ascribes the term “ontology” to nearly everything that is at least a finite controlled vocabulary with unambiguous interpretation of classes and term relationships and with strict hierarchical subclass relationships between classes. We disagree with this overly liberal terminological practice. Ontologies that meet more elaborate criteria, and contain a much richer internal structure were dubbed “heavyweight” and differentiated from so-called “lightweight” ontologies [25]. Among the criteria mentioned for “heavyweight ontologies” are, besides the subtype relation discussed above, also the presence of properties, value

restrictions, general logical constraints, and disjoints. The ACGT MO has been designed, in this respect, to be to a heavyweight ontology. A basic principle of ontology development is that ontologies include only classes (types, universals) but not instances (tokens). Hence the ACGT MO does not include representations of real world instances but only of universals. One of the gold standards to be followed in order to ensure a proper structure of the taxonomy of universals, is the use of a formal subtype relation and the avoidance of the informal *is_a* relations mentioned above. The subtype relation (*is_a*) is formally defined as follows: A *is_a* B if and only if all instances of A are also instances of B.

In general, we embrace the thesis that a properly constructed ontology should steer clear of a taxonomical tree that allows multiple parent classes for the same child class (i.e. one child that inherits from multiple parents). The central aim is to avoid polysemy that often results from multiple inheritances. In the ACGT MO we completely avoided multiple inheritance.

Another problematic case that can be found in a number of medical databases, terminologies and even "ontologies," is the presence of so called *Not Otherwise Specified* (NOS) classes, e.g. "Brain Injury Not Otherwise Specified" or classes like "*UnknownX*" ("*UnknownAffiliation*"). Only recently havea number of revisions of SNOMED CT [26, 27] led to the deactivation of concepts involving the qualifier NOS such as 262686008 Brain injury NOS (disorder) and 162035000 Indigestion symptom NOS (finding). This demonstrates an increasing realist orientation in SNOMED CT. Already Cimino in his famous "Desiderata" essay [28] had counseled against the use of NOS and similar qualifiers. "Universals" of this kind do not, in fact, have any instances of their own; rather, they merely hint at a lack of data or knowledge. The alleged instances of those universals do not exhibit any shared properties, at least not necessarily. Therefore, we avoided such classes in the ACGT MO. The review of pre-existing biomedical ontologies targeting the ACGT domain led to the decision to re-use the FMA and the GO. Furthermore, some existing medical classifications and/or controlled vocabularies have been, or will be, slightly modified and added to the ontology. An example of this type is the TNM system [19].

4 Database Integration Process

The ACGT Semantic Mediation Layer (ACGT-SM) comprises a set of tools and resources that work together to serve processes of Database Integration and Semantic Mediation. The ACGT-MO is a core resource of this system, acting as Global Schema – i.e. a global framework for semantic homogenization – providing the formalization of the domain knowledge needed to support a variety of applications oriented towards clinical research and patient care. The ACGT-SM follows a Local-as-View Query Translation approach in order to cope with the problem of database integration. This means that data is not actually integrated, but is made accessible to users via a virtual repository. This repository represents the integration of the underlying databases, and the ACGT-MO acts here as database schema, providing resources for formulation of possible queries. The virtual repository has the shape of an RDF database, and the language selected for performing queries is SPARQL [29].

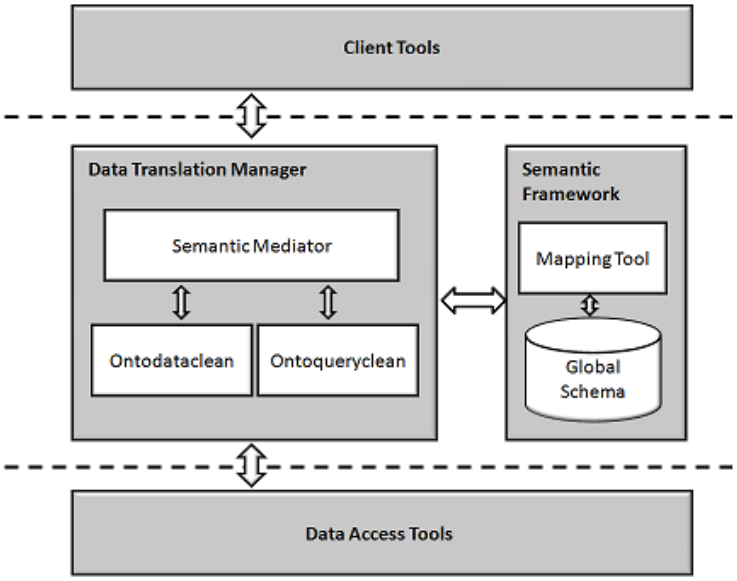

Figure 1: ACGT Semantic Mediator Layer architecture

The ACGT-SM comprises different tools addressing different problems, such as schema level heterogeneities and instance level conflicts both at the query and data levels. These tools are designed as independent web services that collaborate in the mediation process and are coordinated bythe Semantic Mediator. The system also includes a tool devoted to aid in the process of building mappings. This mapping tool uses the ACGT-MO, and is based on a graphical visualization of its structure. The users of this mapping tool navigate the ACGT-MO and an underlying database schema in order to mark the entities that are semantically equivalent. The architecture of this system is shown in Figure 1.

The ACGT-SM exposes its data services using an OGSA-DAI [30] web based interface. The OGSA-DAI middleware allows easy access and integration of data via the grid. However, no grid infrastructure is needed to access these services. The ACGT-SM offers two main services, namely 1) to launch a query, and 2) to browse the schema. The latter shows a subset of the ACGT-SM underlying RDF schema. This subset is built taking into consideration the user profile.

This software system has been tested with clinical relational and image databases [31], obtaining promising results. Currently the consortium is developing a final user query tool, with the aim of helping non technical users in the processes of building and launching queries.

5 Exploitation of the ontology in a clinical trial management system

The integration of existing data sources via the mediator is the general policy of the ACGT project. Yet the ultimate goal of ontology-based information management is to enable the direct integration of semantically consistent data created in different environments (e.g. clinical research, laboratory data, public health data). ACGT aims to provide solutions that demonstrate the possibility of creating data in an ontology-governed way. To explore this approach, an Ontology-based Trial Management System (ObTiMA) is under development that enables those who undertake clinical trials to set up patient data management systems with comprehensive metadata by using the ACGT-MO [32]. This allows seamless integration of data collected in these systems into the ACGT mediator architecture. The main components of ObTiMA are the Trial Builder and the patient data management system. The Trial Builder allows a trial leader to define the master protocol, the Case Report Forms (CRFs) and the treatment plan for the trial in a way that is both semantically compliant with the ACGT MO and user-friendly. From these definitions, the patient data management system can be set up automatically. The data collected in the trial is stored in trial databases whose comprehensive metadata has been rendered from the start in terms of the ACGT-MO. The data can thus be seamlessly integrated through OGSA-DAI services [30] into the mediator architecture. Trial databases with comprehensive ontological metadata and the OGSA-DAI services are both automatically set up from the definitions made by the trial chairman in the Trial Builder

In the following, we briefly describe how the Trial Builder allows the clinician to define all information needed to make integration possible. In setting up a trial, clinicians want to focus on the user interfaces and to adapt them to the specific workflow of the clinical trial planned. They do not wish to be concerned with theoretical aspects and design principles of databases or ontological metadata. Therefore, in ObTiMA, the trial leader defines both, by creating the CRFs for the intended trials. He is assisted by ObTiMA in defining the questions on the CRFs, the order in which the questions will be queried, and constraints on the answer possibilities. Creating a question on the CRF is supported by simply selecting appropriate terms from the ACGT MO. For example, assuming that the clinician wants to collect all information on a patient's gender. He observes that a relation between the classes "Patient" and "Gender" exists in ACGT MO. In creating the corresponding question, he simply has to choose the class *Gender*. The attributes required in order to create the question on the CRF are then determined very easily. E.g. as answer possibilities for the question the values *Male*, *Female*, and *AmbiguousGender* are suggested, because the class *Gender* is defined as an enumeration in the ontology containing these values and a multiple choice question is subsequently automatically created on the CRF.

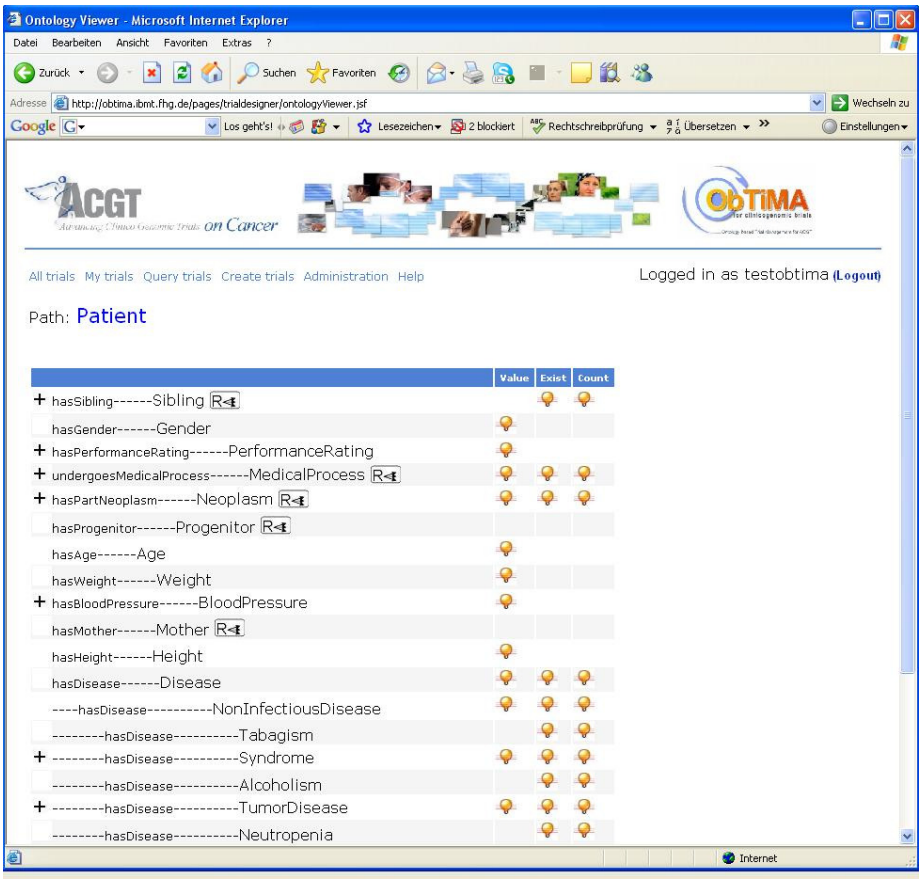

Figure 2: The ObTiMA ontology viewer.

This procedure implements the semantics of the ontology in the CRFs in an automatic fashion.

With the aim of setting up the appropriate database for storing the data, the following attributes are needed for each question: the question itself, the data type of the answer and optionally possible data values, range constraints and measurement units. These attributes will as far as possible be determined automatically from the path the trial leader has selected, but can later be changed according to need and experience of what works best. This process leads to the possibility of lessons learned in integration of the data collected in the clinical trial at hand to be incorporated into the semantics of the ontology. In this way, the ontology itself improves in reflection of advances made by the researchers using it. Through the integration of the ACGT-MO into ObTiMA, data sharing between clinical trials becomes possible. This is necessary to leverage the collected data for further research for example in the creation of cross-trial meta-analyses.

We are aware that this ambitious enterprise requires tools to overcome the gap between clinical practice and biomedical reality representation. Even if an ontology provides natural language definitions for its entities and relationships (in order to

make them human understandable) they are still defined in a way that is not based on practical or clinical perceptions of reality. In order to meet this desideratum, the Trial Builder provides an application-specific view on the ontology, a view that is meant to assist clinicians engaged in clinical practice or clinical trial management.

Recent studies showed that, under three different scenarios, the accuracy of SNOMED coding is only slightly over 50 % [33, 34]. One additional potential advantage of ObTiMA is that it may help put an end to some of the problems currently faced by those using coding techniques to map clinical data unto biomedical terminologies.

Conclusions

The ACGT project provides a novel terminological resource for cancer research and management. It has long been recognized that an obvious application for an ontology resource is to provide a stable common schema for a mediation system such as the one that serves integration across the ACGT network. ACGT has addressed also another problem which is to provide more efficient and reliable tools for coding of clinical data by providing an ontology-driven Clinical Trial Management system which aids the clinician in collecting the data in a way compliant with the ontology.

References

[1] K. H. Buetow, "Cyberinfrastructure: Empowering a 'Third Way,'" *Biomedical Research", Science Magazine*, vol. 308, no. 5723, pp. 821-824, 2005.

[2] M. Tsiknakis, M. Brochhausen, J. Nabrzyski, J. Pucaski, G. Potamias, C. Desmedt and D. Kafetzopoulos, "A semantic grid infrastructure enabling integrated access and analysis of multilevel biomedical data in support of post-genomic clinical trials on Cancer", *IEEE Transactions on Information Technology in Biomedicine,* Special issue on Bio-Grids, March 2008, Vol. 12, No. 2, pp. 205-217.

[3] A. Cali, "Reasoning in Data Integration Systems: Why LAV and GAV Are Siblings," *Proceedings ISMIS 2003, Lecture Notes in Computer Science 2871*, London, Springer 2003, pp 562-571.

[4] http://www.ifomis.org/acgt/1.0

[5] M. Tsiknakis et al., Building a European Biomedical Grid on Cancer, Challenges and Opportunities of HealthGrids: Procs. of the HealthGrid 2006 conference, pp. 247-258, Valencia, Spain, 2006.

[6] http://nciterms.nci.nih.gov/NCIBrowser/Dictionary.do.

[7] http://obofoundry.org/ro.

[8] W. Ceusters, B. Smith, L. Goldberg, "A Terminological and Ontological Analysis of the NCI Thesaurus," *Methods of Information in Medicine* 44:213-220, 2005.

[9] http://www.mged.org.

[10] http://sig.biostr.washington.edu/projects/fm.

[11] http://www.geneontology.org.

[12] B. Smith, M. Brochhausen, "Establishing and Harmonizing Ontologies in an Interdisciplinary Health Care and Clinical Research Environment," in B. Blobel, P. Pharow, M. Nerlich, eds. "eHealth: Combining Health Telematics, Telemedicine,

Biomedical Engineering and Bioinformatics to the Edge, IOS Press, Amsterdam, 2008, pp: 219-234.
[13] M. Dörr. "The CIDOC CRM - An Ontological Approach to Semantic Interoperability of Metadata," *AI Magazine, 24(3*).
[14] http://protege.stanford.edu.
[15] http://www.w3.org/2004/OWL.
[16] B. Smith, W. Ceusters, B. Klagges, J. Kohler, A. Kumar, J. Lomax, C. J. Mungall, F. Neuhaus, A. Rector, C. Rosse, "Relations in Biomedical Ontologies," *Genome Biology*, 6:R46, 2005.
[17] V.T. DeVita Jr., S. Hellman, S.A. Rosenberg (eds.), "Cancer. Principles and Practice of Oncology," 6th Edition, Philadelphia, Lippincott Williams & Wilkins, 2001.
[18] W.A. Schulz, „Molecular Biology of Human Cancers," Dordrecht, Springer, 2005.
[19] C. Wittekind, H.J. Meyer, F. Bootz, "TNM, Klassifikation maligner Tumoren," Berlin, Springer, 2002.
[20] http://suo.ieee.org.
[21] http://www.ifomis.org/bfo.
[22] P. Grenon, B. Smith, L. Goldberg, "Biodynamic Ontology: Applying BFO in the Biomedical Domain." in: *Ontologies in Medicine*, D. M. Pisanelli, Ed., Amsterdam: IOS Press, 2004, pp. 20-38.
[23] A. L. Rector, P. E. Zanstra, W. D. Solomon, J.E. Rogers, R. Baud et al., Reconciling Users Needs and Formal Requirements: Issues in developing Re-Usable Ontology for Medicine, IEEE Transactions on Information Technology in BioMedicine 2(4), pp. 229-242, 1999.
[24] O. Lassila, D. McGuinness, "The role of frame-based representation on the Semantic Web," *Technical Report KSL- 01-02*, Knowledge System Laborartory. Stanford University, Stanford, 2001.
[25] A. Gómez-Pérez, M. Fernández-López, O. Corcho, *Ontological Engineering*. London, Springer, 2004.
[26] http://www.snomed.org/snomedct.
[27] W. Ceusters, K. A. Spackman, B. Smith, "Would SOMED CT benefit from Realism-Based Ontology Evolution?" in *American Medical Informatics Association 2007 Annual Symposium Proceedings, Biomedical and Health Informatics: From Foundations to Applications to Policy,* Chicago, IL, pp. 105-109, 2007.
[28] J. J. Cimino "Desiderata for controlled medical vocabularies in the Twenty-First Century," *Methods Inf Med*; 37(4–5), pp. 394–403, 1998.
[29] SPARQL Query Language for RDF. Available at: http://www.w3.org/TR/rdf-sparql-query.
[30] The OGSA-DAI Project. Available at: http://www.ogsadai.org.uk.
[31] L. Martín, E. Bonsma, A. Anguita, J. Vrijnsen, M. García-Remesal, J. Crespo, M. Tsiknakis, V. Maojo, "Data Access and Management in ACGT: Tools to Solve Syntactic and Semantic Heterogeneities Between Clinical and Image Databases", in Advances in Conceptual Modeling – Foundations and Applications, Lecture Notes in Computer Science, 2007, pp. 24-33.
[32] G. Weiler, M. Brochhausen, N. Graf, A. Hoppe, F. Schera, S. Kiefer, Ontology Based Data Management Systems for post-genomic clinical Trials within an European Grid Infrastructure for Cancer Research, In proc of the 29th Annual International Conference of the IEEE EMBS, Lyon, France, August 23-26, 2007, pp. 6434-6437.
[33] J.E. Andrews, R.L. Richesson, J. Krischer, "Variation of SNOMED CT coding of clinical research concepts among coding experts," *j Am Med Inform Assoc* 2007, 14, 4, p. 497-506.
[34] M.F. Chiang, J.C. Hwang, A.C. Yu, D.S. Casper, J.J. Cimino, J. Starren, "Reliability of SNOMED-CT coding by three physicians using two terminology browsers," *AMIA 2006 Symposium Proceedings*, 2006, p. 131-135.