The ACGT Master Ontology on Cancer – a New Terminology Source for Oncological Practice

Mathias Brochhausen
Institute of Formal Ontology and Medical Information Science (IFOMIS), Saarland University, Saarbrücken, Germany
mathias.brochhausen@ifomis.uni-saarland.de

Gabriele Weiler
Fraunhofer Institute for Biomedical Engineering, Ensheimer Str. 48, 66383 St. Ingbert, Germany
Gabriele.Weiler@ibmt.fraunhofer.de

Cristian Cocos
IFOMIS, Saarland University, Saarbrücken, Germany

Holger Stenzhorn
Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Germany

Norbert Graf
Paediatric Haematology and Oncology, Saarland University Hospital, Homburg, Germany

Martin Dörr
Foundation for Research and Technology-Hellas (FORTH), Institute of Computer Science, Heraklion, Crete, Greece

Manolis Tsiknakis
FORTH, Institute of Computer Science, Heraklion, Crete, Greece
tsiknaki@ics.forth.gr

Abstract

We present a new source of terminology for transnational data exchange in oncology, emphasizing the integration of both clinical and molecular data. In order to achieve best results in semantic interoperability, the ACGT project provides an ontology on cancer research and management. Besides examining pre-existing sources of terminology, we review methods of ontology development, and present best practices to be employed in the development of the ACGT Master Ontology. The clinical trial management system that is currently developed within ACGT constitutes a central use of the ontology at this point.

1. Introduction

The amount of data on cancers and their treatment has exploded over the past years, 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 amount of multilevel datasets (from the molecular to the organ and individual levels) are becoming available to the 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 therefore, confines patient benefit.

The objective of the Advancing Clinico-Genomic Trials on Cancer (ACGT) integrated project is to develop a semantic grid infrastructure in support of multi-centric, post-genomic clinical trials, and thus enable the smooth and prompt transferring 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 developed an ontology (the ACGT Master Ontology (MO) [2]), to be utilized in the context of the selected Local-As-View (LAV) data integration strategy [3]. In such an integration strategy the Master 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 covering 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]. A detailed description of the overall architecture defined by the project and in specific the semantic data integration architecture can be found in [5]. Briefly stated, in a mediator system, queries are formulated in terms of the global schema, translated automatically to the local schemata and then submitted. The answer sets returned from the local systems are translated back to fit the global schema and to be integrated, as, for instance, in order to process cross-system joins.

To integrate a data source into the mediation architecture, a mapping of the local-to-global schema (i.e. the ontology) is required. The ACGT project is developing a user-friendly graphical tool to assist users in performing this mapping in a semi-automated fashion; undertaking the mapping remains, nevertheless, an error-prone and tedious task that has to be supervised by human users [6].

2. The scope of the ACGT MO

The scope of the ontology is determined by the scope of the ACGT project; the latter focuses on the delivery of a knowledge discovery and management platform in the context of post-genomic clinical studies, genomic research, and clinical cancer management and care. The breadth of project scope, that includes numerous different types of data with different security levels, is one of the challenges of the ACGT project, especially in the phase of the ontology development. Hence, the ACGT MO could easily be regarded as a representation of multiple domains. We nevertheless speak of the ACGT MO as a *domain* ontology, in the sense that it represents the entire reality in the domain covered by the ACGT project in a uniform way.

Another challenge of the ACGT MO is to represent clinical reality of cancer management in a highly accurate and consistent way. This requires that clinical reality be the basis of the representation, and that the result proves highly usable in computer applications, like, e.g., the ACGT environment. For that reason, the ontology development team pursued the goal of active and extensive interaction with all of the clinical partners in the project.

3. The ACGT MO

3.1. Technical Details of the ACGT MO

The intention of the ACGT MO 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 in the internet consists of 1300 classes. The ontology was built, and is being maintained, using the Protégé-OWL free open-source ontology editor [7]. It is written in OWL-DL [8] 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. OBO Relation Ontology (RO) [9] has been used as a basis in this regard, as RO has been specifically developed to account for relations in biomedical ontologies [10]. Some properties of scientific observation were taken from the CIDOC CRM [29].

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.

An *is_a* hierarchy of concepts may put things such as a kind of processes, and the kinds of objects it refers to 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 typical concepts appearing together in a particular clinical context, while the full ontology was running behind the scenes. The necessary associations may be found and activated by tracking the workflows common in 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 [11] 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 [12]. Among the criteria mentioned for "heavyweight ontologies" are, besides the subtype relation discussed above, the presence of properties, value restrictions, general logical constraints, and disjoints. The ACGT MO has been designed, in this respect, to amount 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 real world instances but only 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 relations and the avoidance of "informal *is_a*" relations. 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 belief 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 inheritances.

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 types like "*UnknownX*" ("*UnknownAffiliation*"). Only recently, a number of revisions of SNOMED CT [13, 14] 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 [15] had counseled against the use of this and similar qualifiers.

"Universals" of this kind do not, in fact, have any instances; they merely hint to 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 [16].

3.3. The upper level: Basic Formal Ontology (BFO)

In order to provide a consistent and sound representation, the ACGT MO employs the resources of a Top Level Ontology or Upper Level Ontology, which is, according to the Standard Upper Level Ontology Working Group of IEEE, "limited to concepts that are meta, generic, abstract and philosophical, and therefore are general enough to address (at a high level) a broad range of domain areas. Concepts specific to given domains will not be included; however, this standard provides a structure and a set of general concepts upon which domain ontologies could be constructed" [17]. We haven chosen the Basic Formal Ontology (BFO) [18] as Top Level for the ACGT MO since BFO has proven to be highly applicable to the biomedical domain. BFO rests on four principles which also governed the development of ACGT MO [19]: *Realism* is defined as the view according to which reality and its constituents exist independently of our (linguistic, conceptual, theoretical, cultural) representations. *Perspectivalism* involves the recognition that reality can be captured in many different, though equally good, representations (good in the sense of being true). *Fallibilism* involves commitment to the idea that it may be the case that portions of our current knowledge are incorrect hence our current purported reality representations are not representations after all. *Adequatism* is the position that

a good theory of reality must do justice to all of the different phenomena that reality contains. It is opposite to reductionism.

We believe that these principles are a crucial part of the attitude one has to adopt regarding the development of a reality-based ontology.

4. 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 state of the art 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 considerably 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, as we do in ACGT, the National Cancer Institute Thesaurus (NCIT) is a terminology resource of obvious relevance [20]. 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 meagre for our purposes, with only a fraction of NCIT terms being supplied with the formal definitions of the sort required by its official description logic 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 [9]. Further, the NCIT is marred by a number of problems in its internal structure and coverage [21], including problems in the treatment of *is_a*. For a quick illustration of the inadequate treatment of *is_a* in the NCIT let us consider the NCIT class *Organism*, which includes among its subtypes *OtherOrganismGroupings*;' with this we have *OtherOrganismGroupings is_a Organism* [20]. Given the formal definition of the subtype relation this is clearly wrong, as 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 [22]. Yet, even this highly used resource shows considerable deficiencies, like informal *is_a* relations. The inconsistency becomes obvious when the textual definitions--which are an asset of MGED--are taken into account: According to the MGED ontology *Host* is a subclass to *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)" [22]. 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" [22]. That an organism or an organism's part is a description is clearly a crude category mistake.

Nevertheless, this does not preclude that, for some aspect of the ACGT domain, 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) [23] and the Gene Ontology (GO) [24] since they both fulfill the requirements on coherence and theoretical rigor specified in section 3.

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 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, [25]) 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, to the extent to which they were not already present in BFO.

To conclude, regarding the key issue of the ACGT domain, developing a new ontology was imperative, since no single ontology or set of ontologies had the respective coverage and logical consistency.

5. Quality assessment in ontology development

Even though there are means to classify ontologies with regard to their complexity and amount of information represented in them, ontology development needs to appeal to standards of quality. Basic criteria for ontology assessment can be derived from marking the differences between ontologies and their terminological predecessors. The following assets are typical for ontologies: a logical structure to support algorithmic processing, a concern for the reality to

which the terms in an ontology relate, the possibility of interoperability of ontologies developed for the representations of related domains of entities, and a coherent strategy for quality assurance, based on user feedback and empirical testing, for update and maintenance in light of scientific advance and for evolutionary improvement of the ontology as a whole.

We maintain that a core aspect of ontology assessment consists in establishing whether these components are indeed observed and complied with. Furthermore, terminological and theoretical conventions targeting the distinctions between reality, mental representation and intersubjective representation (e.g. ontologies) [26] must be respected.

This aspect is addressed by aligning the ACGT MO with the principles of the OBO Foundry (e.g. the ontology is expressed in a common formal language, has a clearly specified and clearly delineated content, and is openly available), which introduces "a new paradigm for biomedical ontology development by the establishment of gold standard reference ontologies for individual domains of inquiry" [27]. Our objective is for the ACGT MO to become a member of the OBO Foundry once it is fully completed. We, further, contend that all ontologies used by the mediator should also subscribe to the standard of OBO Foundry. This is already the case for FMA and GO.

6. 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, etc.). 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 in terms of the ACGT-MO [28]. 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 chairman 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. This collected data is stored in trial databases whose comprehensive metadata has been rendered in terms of the ACGT-MO. The data can thus be seamlessly integrated through OGSA-DAI services [29] 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 try to integrate and adapt them into the specific workflow of the clinical trial planned. They should not be concerned with theoretical aspects and design principles of databases or ontological metadata. Therefore, in ObTiMA, the trial chairman defines both, by creating the CRFs for his trials. He is assisted 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 concepts from the ACGT MO. For example, assuming that the clinician wants to collect all information on the patient's gender. He observes that a relation between the classes "Patient" and "Gender" does exist. In creating the corresponding question, he simply has to choose the class *Gender*. The attributes required in order to create the possible answers on the CRF are then determined automatically. The values allowed are set automatically to *Male*, *Female*, and *AmbiguousGender* since 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.

This procedure implements the semantics of the ontology in the CRFs in an automatic fashion. We expect that the description alluded to above be a path from the ontology starting at the class *Patient*, as this is normally the focal point of CRFs.

With the aim of setting up the appropriate database for storing the data, the following attributes are needed for each question: the question itself, data type of the answer and optionally possible data values, range constraints and measurement units. These attributes will, as much as possible, be determined automatically from the path the trial chairman has selected, but can later be changed by the trial chairman to a certain extent. This process leads to the direct integration of the data collected in the clinical trial at hand into the semantics of the ontology. 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 like cross-trial meta-analysis.

We are, nevertheless, 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 (is, in other words, '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 in clinical practice, as well as when tackling workflows typical of clinical trial management [30].

7. Conclusions

The ACGT project provides a novel terminological source for cancer research and management. The main goal was to create a "heavyweight" ontology enabling semantic data integration. The main characteristic of the development process was that clinical expertise was brought together with best practice in ontology development.

This is an extremely important point, since the review of existing terminologies and ontologies in cancer-focused biomedicine revealed that most such sources fail to be coherent due to a lack of logical and theoretical principles respected in the development. Dealing, on the other hand, with representations of ontological nature can be hard and time consuming for end-users, e.g. clinicians. Ontology development and ontological engineering must overcome these problems and must subscribe to a truly interdisciplinary practice.

8. References

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

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

[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] 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.

[5] 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 (accepted for publication – available at http://ieeexplore.ieee.org/xpl/tocpreprint.jsp?isnumber=4358869&Submit3=View+Articles&punumber=4233).

[6] Y. Kalfoglou, M. Schorlemmer, Ontology mapping: the state of the art, The Knowledge Engineering Review, 18(1) pp. 1-31, 2003.

[7] http://protege.stanford.edu.

[8] http://www.w3.org/2004/OWL.

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

[10] 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.

[11] 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.

[12] A. Gómez-Pérez, M. Fernández-López, O. Corcho, *Ontological Engineering*. London, Springer, 2004.

[13] http://www.snomed.org/snomedct.

[14] 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.

[15] J. J. Cimino "Desiderata for controlled medical vocabularies in the Twenty-First Century," *Methods Inf Med*; 37(4–5), pp. 394–403, 1998.

[16] C. Wittekind, H. J. Meyer, F. Bootz, *TNM. Klassifikation maligner Tumoren*. 6. Aufl., Heidelberg, Springer, 2005.

[17] http://suo.ieee.org.

[18] http://www.ifomis.org/bfo.

[19] 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.

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

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

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

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

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

[25] M. Dörr. "The CIDOC CRM - An Ontological Approach to Semantic Interoperability of Metadata," *AI Magazine, 24(3*).

[26] B. Smith, W.Kusnierczyk, D. Schober, W. Ceusters, "Towards a Reference Terminology for Ontological Research and Development in the Biomedical Domain", KR-MED, 2006.

[27] http://obofoundry.org.

[28] 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.

[29] www.ogsadai.org.uk.