Design Principles of the ACGT Master Ontology: Examples and Discussion

Project Number: FP6-2005-IST-026996
Deliverable id: D 7.7
Deliverable name: Design Principles of the ACGT Master Ontology: Examples and Discussion
Submission Date: December 2008

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 7.7
Document name:	Design Principles of the ACGT Master Ontology: Examples and Discussion
Document type (PU, INT, RE)	INT
Version:	3.0
Submission date:	23/11/2008
Editor: Organisation: Email:	Cristian Cocos IFOMIS, USAAR cristi@ieee.org

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

ABSTRACT: The present deliverable details and discusses the major design decisions underlying the development of the ACGT Master Ontology. Examples and counterexamples are being given, and alternatives are being examined. It is also intended as a guide for future developers of the ontology, as well as the backbone for a manual on general ontology development in its own right.

KEYWORD LIST: Ontology, Ontology Development, Design Principles

MODIFICATION CONTROL			
Version	Date	Status	Author
1.0	28/10/08	Draft	C. Cocos
2.0	15/11/08	Draft	C. Cocos
3.0	20/11/08	Final	C. Cocos

List of Contributors

- Cristian Cocos, USAAR
- Mathias Brochhausen, USAAR
- Erwin Bonsma, Philips
- Luis Martin, UPM

Contents

Executive Summary

Chapter 3 elaborates a distinction between two types of ontologies that has been found to be relevant in the present context. Chapter 4 constitutes the core of this document, as it discusses the major design choices that ACGT MO developers have adopted. Chapter 5 states the multigranular character of the ontology, while chapter 6 lists its sources. The methodology adopted in building the ontology has been detailed in chapter 7. Finally, chapter 8 gives examples of SPARQL queries formulated in ACGT MO terms.

1 Introduction

Purpose of this document

This deliverable aims at familiarizing ACGT MO users with its underlying design philosophy via diagrams and examples. Its ultimate goal is to constitute the core of the MO documentation endeavours, hence enabling future MO developers to continue development efforts towards achieving MO's targets by offering deep and detailed insight into its founders' ontological attitude and viewpoint since MO's very inception.

Introduction

The discipline of knowledge engineering grew out of the early work on expert systems in the seventies. With the growing popularity of knowledge-based systems (as these were by then called), there also arose a need for a systematic approach to building such systems, similar to methodologies in mainstream software engineering. Over the years, the discipline of knowledge engineering has evolved into the development of theory, methods and tools for developing knowledge-intensive applications. In other words, it provides guidance about when and how to apply particular knowledge-presentation techniques for solving particular problems.

The early expert systems were based on an architecture which separated domain knowledge, in the form a knowledge base of rules, from a general reasoning mechanism. This distinction still is still valid in knowledge engineering practice. In the early eighties a number of key papers were published that set the scene for a systematic approach to knowledge engineering.

In the nineties the attention of the knowledge-engineering shifted gradually to domain knowledge, in particular reusable representations in the form of ontologies. A key paper, which also quite wide attention outside the knowledge-engineering community was Gruber's paper on portable ontologies [Gruber 93]. During this decade ontologies were getting widespread attention as vehicles for sharing concepts within a distributed community such as the web. Gruber defines an ontology as an "explicit specification of a conceptualization". Several authors have made small adaptations to this. A common definition nowadays is: *An ontology is an explicit specification of a shared conceptualization that holds in a particular context.* The addition of the adjective "shared" is important, as the primary goal of ontologies in computer science was to enable knowledge sharing. Up until the end of the nineties "ontology" was a niche term, used by a few researchers in the knowledge engineering and representation field. The term is now in widespread use, mainly due to enormous need for shared concepts in the distributed world of the web. People and programs need to share at least some minimal common vocabulary. Ontologies have become in particular popular in the context of the Semantic Web effort.

2 Types of Ontologies

For our limited purposes, a useful division of ontologies is the one between the following two types: (i) formal ontologies (aka *foundational* ontologies), and (ii) domain-specific (aka *material*) ontologies [Spear 06].

Formal ontologies. Foundational ontologies stay closest to the original philosophical idea of "ontology." These ontologies aim to provide conceptualizations of general notions, such as time, space, events and processes. Some groups have published integrated collections of foundational ontologies. Two noteworthy examples are the SUMO (Suggested Upper Merged Ontology) and DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering). The former, for example, uses the term "upper ontology" in this respect: "An upper ontology is 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 [i.e. SUMO] will provide a structure and a set of general concepts upon which domain ontologies (e.g. medical, financial, engineering, etc.) could be constructed" [http://suo.ieee.org/]. Ontologies for part–whole relations have been an important area of study. Unlike the subsumption relation (*is_a*), the cornerstone of any ontology, part–whole relations are usually not part of the basic expressivity of the representation language. In domains dealing with large structures, such as biomedicine, part–whole relations are often of prime importance. A simple baseline representation of part–whole relations is given by [Rector and Welty 05]. Our favorite supplier of relational resources (part-whole and others) is the Relation Ontology (RO) [Smith 05], and will be briefly discussed below.

Lexical resources such as WordNet, can also be seen as foundational ontologies, although with a weaker semantic structure. WordNet defines a semantic network with 17 different relation types between concepts used in natural language. Researchers in this area are currently proposing richer semantic structuring for WordNet. The original Princeton WordNet targets the English–American language; Word-Nets now exist or are being developed for almost all major languages.

Domain-specific ontologies. Although foundational ontologies are receiving a lot of attention, the majority of ontologies are domain-specific: they are intended for sharing concepts and relations in a particular area of interest. One domain in which a wide range of ontologies has been published, and at which our present efforts are aimed, is biomedicine. A typical example is the Foundational Model of Anatomy (FMA), which describes some 75,000 anatomical entities. Other well-known biomedical ontologies are the Unified Medical Language System (UMLS), the Simple Bio Upper Ontology, and the Gene Ontology (GO). Domain ontologies vary considerably in terms of the level of formalization. Communities of practice in many domains have published shared sets of concepts in the form of vocabularies and thesauri. Such concept schemes typically have a relatively weak semantic structure, indicating many hierarchical (broader/narrower) relations, which most of the time loosely correspond to subsumption relations. This has triggered a distinction in the ontology literature between weak versus strong ("heavyweight" [Gómez-Pérez 04]) ontologies. The ACGT MO has been specifically designed as to amount to a heavyweight ontology.

3 ACGT MO (Major) Design Choices with Discussion and Illustrations

1. DESIGN CHOICE NUMBER ONE: The adoption of a radically modified definition of the term "ontology," in compliance with the principle of realism (see Glossary).

The following definition of 'ontology' has recently been proposed ([Smith 06]), and contains most of the elements that it will be important to discuss here: *an ontology is a representational artifact whose representational units are intended to designate universals in reality and the relations between them*.

This definition has two parts. The first identifies an ontology as a *representational artifact* consisting of *representational units*, while the second has to do with what the representational units in such an artifact are intended to refer to or be *about*. A short explanation of each term is in order.

The human world is full of representations and representational artifacts. The key feature common to all representations is that they make reference to or are about something else. Thus a representation is an idea, an image or a description that refers to some entity or entities external to itself. The memory that one has of the Alexander Nevsky Cathedral in the Bulgarian capital Sofia, is a representation in one's mind that is about or refers to an entity other than itself, namely the actual Alexander Nevsky Cathedral that exists in Sofia. Similarly, the thoughts of a scientist as he looks through a microscope at bacteria, namely the thoughts that "these are bacteria," are mental representations that, taken together, point beyond themselves and make reference to the actual existing bacteria that are under investigation. It is, indeed, one of the most basic features of human thought that beliefs, desires and experiences in general point beyond themselves and refer to the objects that they are about. However, representations by themselves are not yet ontologies in the sense in which we are here interested. Ontologies have the important further feature of being representational artifacts.

A representational artifact is an entity which makes pre-existing cognitive representations from the minds of its author publicly available. Representational artifacts include things such as signs, books, pictures and diagrams. A key feature of representational artifacts it that they include ledgers or rules for their interpretation. Thus, maps do not simply come color coded, they also come with a key or table that makes it possible to interpret their color coding as representing certain kinds of things (countries, oceans, mountain ranges, etc.), and the words in which these tables and keys are written themselves have publicly available rules for their interpretation as referring to things in the world, namely the semantics of natural language itself. A simple kind of representational artifact would be a picture of the Alexander Nevsky Cathedral in Sofia drawn based on the mental representation that comprises one's memory of having once seen it. Whereas one's memory is a cognitive representation, the picture drawn based on it is a representational artifact intended to refer to the same thing to which the original memory refers to, namely the Alexander Nevsky Cathedral, and yet it exists independently of one's own mind or thoughts in a form that is publicly observable and inspectable.

At this point two important caveats are to be issued:

(a) When constructing a representational artifact for use in science, such as an ontology, based on cognitive representations or concepts in the minds of individual subjects, the goal is *not* to accurately represent in a publicly accessible way the *representations* or *concepts* that exist in the individual's minds, *but* rather the *things in reality* that these representations are representations *of*. (See also the "Realism" entry in the Glossary)

(b) There is a fundamental distinction between *using* such artifacts to make reference to things in reality, the entities that they are representations of, on the one hand, and *mentioning* such artifacts by engaging in discourse about them on the other. The construction of coherent functional ontologies requires that this *use-mention distinction* be strictly taken into account.

Here are a few examples of the latter fallacy drawn from the practice of widely-available biomedical ontologies and controlled vocabularies:

In the Medical Subject Headings (MeSH) one can find the following hierarchical relationship "National Socialism *is_a* MeSH Descriptor." Here National Socialism, which is a kind of political movement that existed in the world, is identified as a kind of term in the MeSH database, however the definition of 'National Socialism' in MeSH as "The doctrines and policies of the Nazis or the National Social German Workers party, which ruled Germany under Adolf Hitler from 1933-1945. These doctrines and policies included racist nationalism, expansionism, and state control of the economy. (from Columbia Encyclopedia, 6th ed. and American Heritage College Dictionary, 3d ed.)" makes it clear that the use and mention of the term 'National Socialism' have here been confused. Similarly, in the National Cancer Institute Thesaurus (NCIT), the following definition of 'Conceptual Entities' can be found "An organizational header for concepts representing mostly abstract entities." As a definition there are a number of problems with this. However, what is important here is simply that, once again, the use and mention of a term have been conflated. Whereas we would expect a database that is about things in reality to provide definitions that would tell us facts about the basic features of those things (in this case the basic features of conceptual entities), what we get here is an explanation of how the *term* 'conceptual entity' is used as a part of the representational artifact that is the NCIT. Once again, statements about things (use) and statements about the words used to mention things (mention) have been conflated. Other examples include the definition of 'mouse' as "name for the species *mus musculus*" in BIRNLex, the entry "Bacterium causes Experimental Model of Disease" in the Unified Medical Language System (UMLS), the definition of 'animal' as "a subtype of Living Subject representing any animal-of-interest to the Personnel Management domain" from the HL7 Glossary, and "living subject is_a code system" from the HL7. All of these are examples of ways that databases and ontologies can go wrong when they fail to keep separate the use of the ontology and its terms to refer to things in reality, and the mention of terms and elements of the ontology in discourse that is explicitly about it, its construction and its constituent elements.

Finally, as a final comment related to design choice number one, we feel at this point that we have to give a quick explanation of the term "universal" used in the (modified) definition of ontology. It is a basic assumption of scientific inquiry that nature is structured, ordered and regular, at least to some degree. Though scientists always perform experiments and make observations regarding particular objects, what they are actually interested in are the generalizations about the structure, order and regularity that exists in nature that such experiments and observations make possible. Universals are *that which is general or abstract in reality*. They are the philosophical explanation of the structure, order and regularity that is to be found in nature, and they are what all members of a natural kind, grouping or species (for example the kind “feline” or “mammal”) have in common. Universals are repeatable in the sense that they *can be instantiated by more than one object and at more than one time*, whereas particulars, such as myself, my right kidney, and specific political administrations, are non-repeatable: they can exist only in one place and during one period of time. Because of this, universals do not have a determinate location in space or time. Rather, they exist at all times and in all places where particular entities instantiating them exist.

As opposed to universals, *particulars* are the individual denizens of reality. Particulars instantiate universals, but cannot themselves be instantiated. It is in virtue of instantiating a given universal that two particulars will be similar in some respect. Particulars exist in space and time, and come into and pass out of existence. It is possible to causally interact with, directly see with one’s eyes, touch and smell particulars, but not universals. For example, the universal *DNA microarray* is an abstract entity that is instantiated by and accounts for the similarities amongst all particular DNA microarrays, but unlike its instances, the universal has never been and cannot be, say, analyzed under a microscope.

2. DESIGN CHOICE NUMBER TWO consists in the adoption of Basic Formal Ontology (BFO) [http://www.ifomis.org/bfo] as upper/formal ontology.

The BFO taxonomy (figure 1) makes use of a basic top-level distinction (“the great divide”) between two kinds of entities: substantial entities or continuants (entities that endure through time while maintaining their identity) on the one hand, and occurrents or perdurants (entities that happen, unfold, or develop in time) on the other. Corresponding to these two kinds of entities are two basic and distinct perspectives that can be taken on the world, neither of which can fully capture or represent the features of reality represented by the other: these are the SNAP and SPAN perspectives or ontologies respectively. Each of these basic perspectives can also be used to represent entities at different levels of granularity/abstraction, resulting in further perspectival subdivisions of the basic SNAP and SPAN ontologies. For our present purposes, suffices to mention that the SNAP ontology recognizes three major categories of continuants: dependent continuants, independent continuants and spatial regions, while SPAN includes processual entities and spatiotemporal regions. We will quickly explain some of these categories (namely those that are not self-explanatory), and illustrate them with examples taken from the ACGT Master Ontology.

Figure 1: BFO (arrows represent *is_a* relations)

The defining feature of *independent* continuants is that they are the kinds of things in which other continuants, such as qualities and dispositions, can inhere. They are the bearers of qualities and other dependent continuants (among which the category of realizable continuant will be exemplified further below). Examples: HumanBeing, Liver, Prostate, InsuranceCompany, Laboratory, Microarray etc.

The defining feature of *dependent* continuants is that they are the kinds of things (qualities, roles, functions) that inhere in or are born by something else (namely independent entities). Examples: the SurfaceSize of a Liver, the Weight of a Prostate, the Gender of a HumanBeing, the Disease of an Organism, the Role of being an Implant, the Function of the Heart to pump blood etc.

Realizable entities are dependent continuants that inhere in continuant entities and are not exhibited in full at every time in which they inhere in an entity or group of entities. The exhibition or actualization of a realizable entity is a particular manifestation, functioning, or process that occurs under certain circumstances. Examples: the Role of being a (Biological) Marker, the Function of the reproductive organs, the Disposition of blood to coagulate, the Disposition of metal to conduct electricity etc. According to the definition, hence, a simple criterion can be employed in order to avoid a frequent mistake by which entities that otherwise belong squarely in the RealizableEntity class receive a different classification (mostly as some IndependentContinuant or other): if one can confidently "subsume" one and the same entity at various stages of its life under one and the same class, it is very likely that the latter belongs on the RealizableEntity branch.

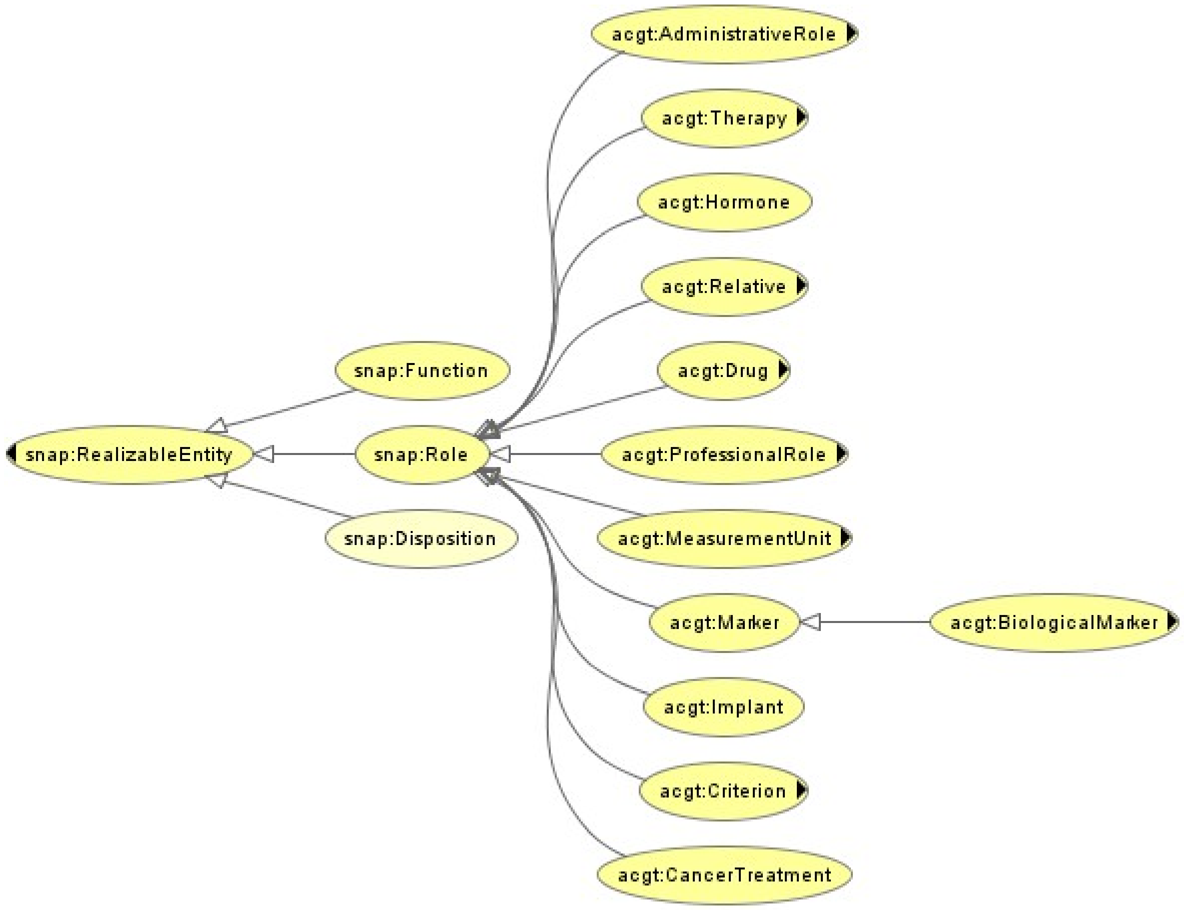


Figure 2: RealizableEntity

A ProcessualEntity is an occurrent entity that exists in time by occurring or happening, has temporal parts, and always depends on some SNAP entity or entities. Examples: the Life of an Organism, the Process of Meiosis, the processes of Biopsy, ClinicalExamination, PregnancyTest, TumorDevelopment, the process of cell division etc.

Finally, by way of *motivation*, our choosing BFO as top-level ontology comes as result of an intense research activity devoted to assessing several well-known and widely utilized ontological frameworks and controlled vocabularies, at both formal/upper and lower (domain-specific) level. This research activity has been concretized in Deliverable 7.1, where it was concluded that: (a) while several candidate formal ontologies may comply with the criteria that outline a heavyweight ontology, none, with the exception of BFO, meet a number of indispensable theoretical conditions, some of which will be mentioned in what follows (admittedly, at this point, a great deal of weight has been placed on ACGT MO developers' personal philosophical preference and ideological bent); (b) positively no existing domain/biomedical ontology or controlled vocabulary fulfil the high ontological standards required in order to secure a firm ontological grounding for the ACGT project.

3. Design Choice Number Three amounts to assenting to BFO's design principles (realism, perspectivalism, fallibilism and adequatism—see Glossary).

4. Design Choice Number Four: enforcing a strict subsumption hierarchy, based on a formally specified *is_a* relation, as opposed to a loose "subclass" hierarchy.

The great majority of currently existing ontologies incorporate relations that connect their terms ("nodes"). Such relations, however, are being used in mostly informal ways, often providing no definitions at all, so that the resulting logical interconnections are far from clear. While the general aspect of relations will preoccupy us further down below, even the basic taxonomical relation, the foundation of any ontology, is not always used in consistent fashion. A *formal is_a* relation should *at the very least* ensure that an instance of a class is also an instance of its parent class, which is not what always happens in the case of loose taxonomies as encountered in many contemporary fashionable ontologies, both formal and domain specific. To take an example from [Lassila 01], "Yahoo […] provides a small number of top-level categories such as apparel and then dresses as a kind of (women's) apparel. […] Its hierarchy is not a strict subclass or 'isa' hierarchy however. This point […] seems to capture many of the 'naturally occurring ontologies' on the web. In these hierarchies it is typically the case that an instance of a more specific class is also an instance of the more general class but that is not enforced 100% of the time. For example, the general category apparel includes a subcategory women (which should more accurately be titled women's apparel) which then includes subcategories accessories and dresses. While it is the case that every instance of a dress is an instance of apparel (and probably an instance of women's dress), it is not the case that a dress is a woman and it is also not the case that a fragrance (an instance of a women's accessory) is an instance of apparel. This mixing of categories such as accessories in web classification schemes is not unique to Yahoo – it appears in many web classification schemes."

5. Design Choice Number Five: Avoid instances/tokens: A basic principle of ontology development is that ontologies include only classes (types, universals—see discussion above) but not instances (tokens). Hence the taxonomic tree of the ACGT MO does not include terms/names for real world instances, but only names for classes/universals.

Figure 3 below exemplifies a hypothetical situation where an individual has been appended to the taxonomy of universals:

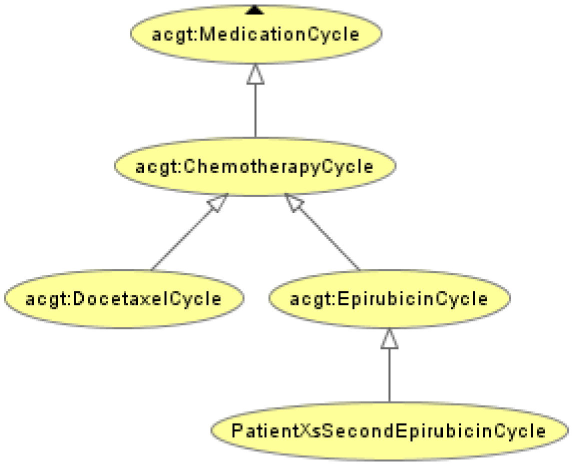


Figure 3: Individual in taxonomic tree

6. DESIGN CHOICE NUMBER SIX: overt avoidance of (non-trivial) multiple inheritance in the hierarchy of universals. 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.

The CIDOC CRM [http://cidoc.ics.forth.gr/] (figure 4) is an example of an ontology that allows multiple inheritance.

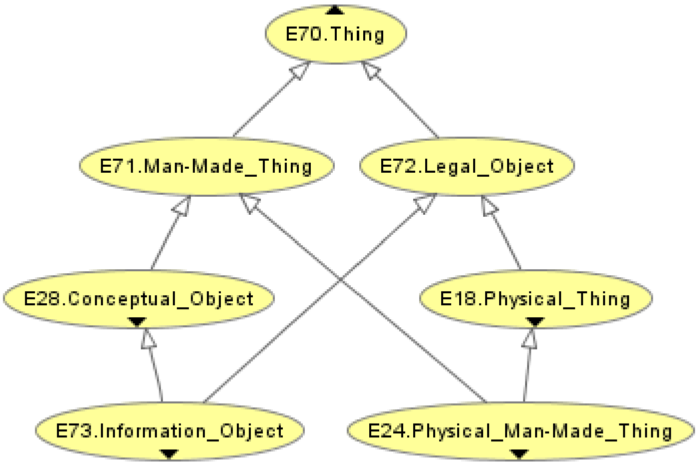


Figure 4: Multiple inheritance in the CIDOC CRM

As it can be seen from the above figure, class E24 (Physical_Man-Made_Thing) inherits both class E71 (Man-Made_Thing) and class E18 (Physical_Thing), and this while there is *no inheritance relation* (*isA*) between E18 and E71. Note that due to the transitivity of the *isA* relation, the existence of an inheritance relation between these two would obviate the "multiple inheritance" claim: E24 (Physical_Man-Made_Thing), for example, is both a Physical_Thing (E18), and a Legal_Object (E72), without multiple inheritance obtaining. Strictly terminologically speaking, we could choose to dub these two inheritance scenarios as *trivial* (i.e. where the relation at stake can be inferred via transitivity of *isA*) versus *non-trivial* (where the relation cannot be so inferred, and has to be asserted), but we have chosen a less verbose way of emphasizing the deviant circumstance via dropping the multiple inheritance appellative in the regular case. The same discussion goes for class E73 (Information_Object), which inherits both E28 (Conceptual_Object) and E72 (Legal_Object). Accordingly, the following relations hold:

Physical_Man-Made_Thing *isA* Man-Made_Thing

Physical_Man-Made_Thing *isA* Physical_Thing

Information_Object *isA* Conceptual_Object

Information_Object *isA* Legal_Object

In the ACGT MO we chose to deal with polysemy by undertaking a disambiguation of naturally-occurring polysemic terms; e.g. Birth in natural language denotes, among others, both the beginning of Life (a ProcessBoundary), and a Process simpliciter—namely the very process of giving birth. The latter can also be encountered in the specialty literature under the more specific term of Parturition (with proper part Labor), which we chose to adopt, while leaving the term Birth under its former, more common, reading (see figure 5).

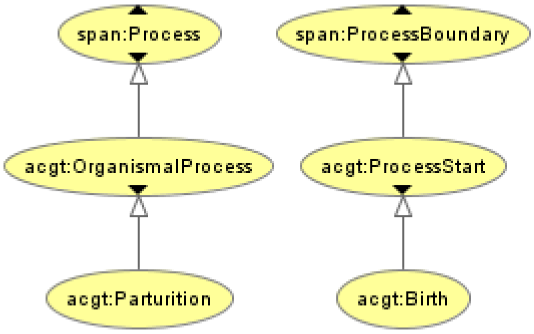


Figure 5: Resolving polysemy

7. Design Choice Number Seven (*completeness desideratum*): the union of a class' children should exhaust the class—that is, the set of children should constitute a *cover* of that class.

8. Design Choice Number Eight: Sibling classes should be disjoint.

This, together with the completeness desideratum (3), should in fact ensure that the set of (first-order) children of a class should constitute a partition of that class.

An important exception to the disjointness rule has, nevertheless, been tolerated in the ACGT MO, due to circumstances relating to the mapping process; more specifically, mapping of SPARQL queries would have been considerably hindered by the existence of OPTIONAL and FILTER blocks—blocks normally required by a definition of the PrimaryTumor class in terms of the non-existence of tumors whose metastasis that primary tumor is. We have, hence, opted to add both the PrimaryTumor and Metastasis classes to the asserted taxonomy, even though this, again, violates design directive number seven (see figure 5): aside from haematooncological tumors, all other tumors (mixed, dysontogenic, neuroendocrine, carcinoma and sarcoma) have both instances that belong in the PrimaryTumor class and the Metastasis class. Note that the two “offending” classes (PrimaryTumor and Metastasis) should ideally be conceived as *roles*.

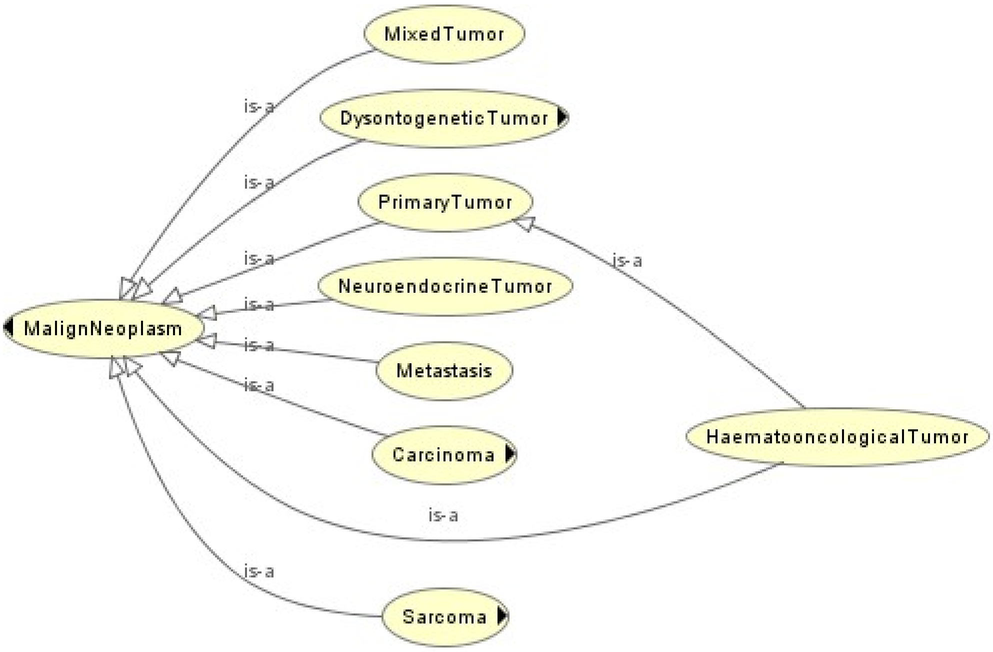


Figure 6: Disjointness violations in the ACGT MO

Note also that as of this writing, the master ontology includes rather few disjointness stipulations, as there is considerable *content*-related debate in this respect; we, however, expect that further versions will make progress towards exhibiting disjoint classes more fully and faithfully.

Prompted by similar considerations, we do not exclude further violations of the disjointness rule in the future, even though we would prefer that the amount of such exceptions be kept as low as possible.

9. DESIGN CHOICE NUMBER NINE: Avoidance of UnknownX and related classes.

A common conduct among developers of medical databases, terminologies, and ontologies, is the inclusion of classes of the UnknownX type ("*UnspecifiedTumorStage*," "*UnknownAffiliation*"). "Universals" like these do not, however, have any instances, but merely hint to a lack of data or knowledge, and hence represent an illegitimate epistemic intrusion in what should otherwise constitute a faithful picture of *reality*, of what there is. The alleged instances of those universals do not exhibit any shared properties, at least not necessarily. While we have striven to avoid the addition of such classes to the ACGT MO, we have seen ourselves compelled to include some via the import of well-known and consecrated medical databases like the TNM [Wittekind 05]. While this import will be discussed later on, we have chosen to illustrate this practice in figure 7 below:

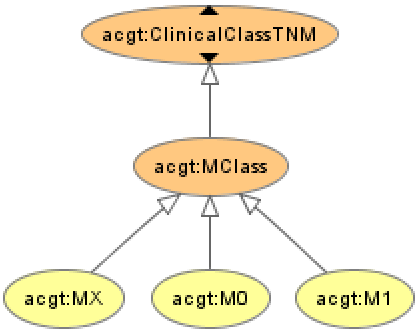


Figure 7: TNM's MX class

Here TNM's MX class reads "Presence of distant metastasis cannot be assessed."

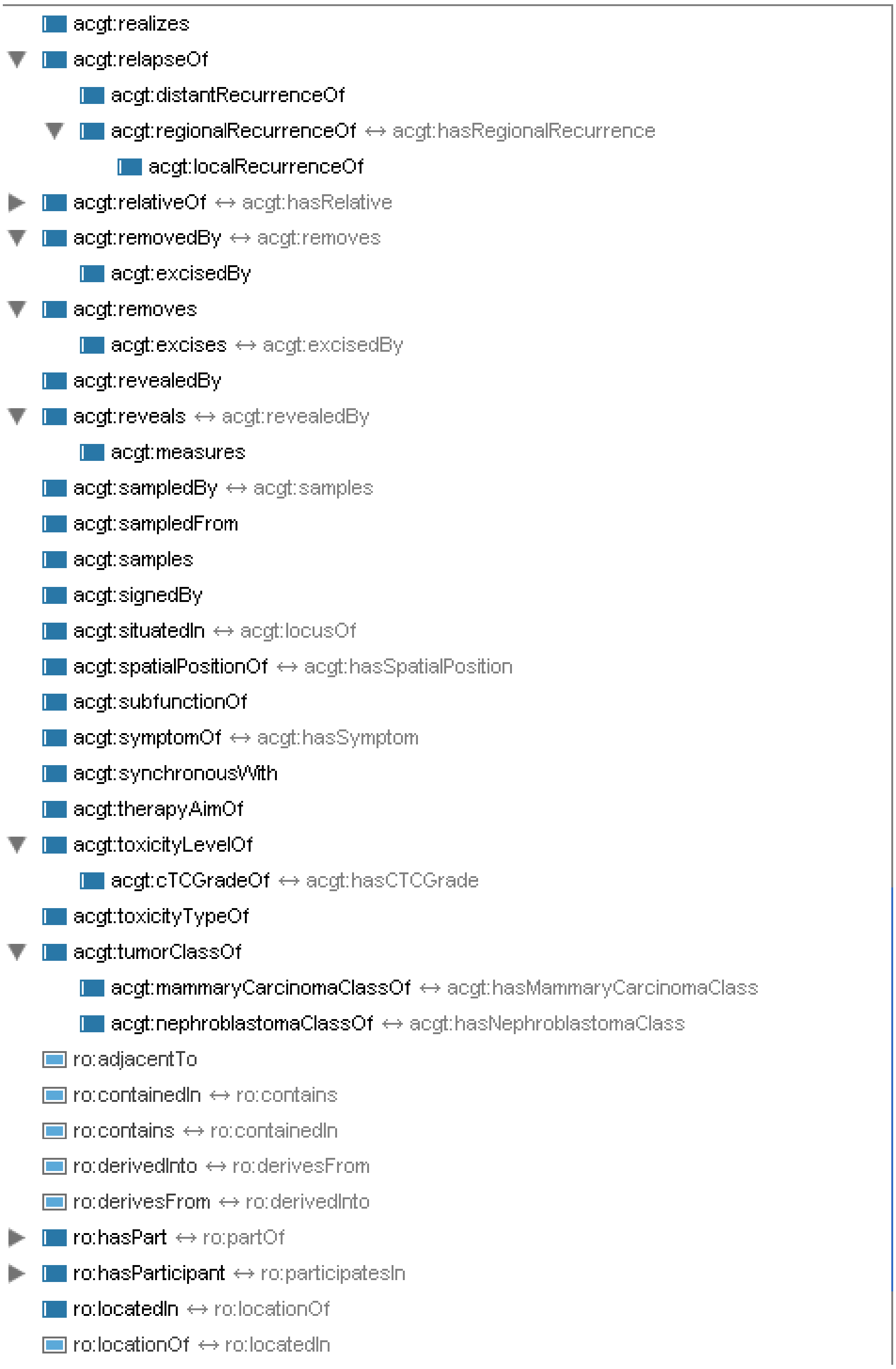
10. DESIGN CHOICE NUMBER TEN: Adoption of Relation Ontology (RO) [Smith 05] as a source of, and insight for new, relations/properties.

The ACGT MO not only represents classes as linked via the basic taxonomical relation ("*is_a*"), but, in good heavyweight ontology style, it connects them via other semantic relations called "properties" in OWL terminology. OBO Relation Ontology (RO) has been used as a basis in this regard, as RO has been specifically developed to account for relations in biomedical ontologies. Moreover, as with the advent of OWL-based ontologies the business of controlled vocabularies and supervised scientific nomenclature has entered a new era, ontology-grounded efforts directed at regimenting widely (but inconsistently) used biomedical terms have blossomed quite spectacularly into a huge internationally coordinated movement, which involves thousands of researchers at the forefront of biomedical science. The OBO Foundry [Smith 07], as it is called, has set as its core mission the establishment of a basic set of ontology design guidelines that are supposed to enable scientists to build interoperable ontologies, hence fostering communication and collaboration in a world that finds itself in increasing danger of being smothered under the pressure of mountains of brute, unanalyzed, quasi-chaotic data. The point of relevance to us at this stage is that using RO for relation regimentation is part of the OBO foundry criteria of ontological excellence. The designers of the ACGT MO have hence set as one of their goals the inclusion of the ACGT MO among OBO Foundry ontologies.

In order to allow for automated maneuverability via reasoning and consistency detection tools, and in order to provide for productive inference and analysis, present days ontologies have to be endowed with expressive powers well beyond what traditional ontologies were about. As mentioned above, traditional ontologies were little more than taxonomies of Existence/Being, hence based on "*is a*" hierarchies (aka subsumption hierarchies). As a knowledge representation style, however, users demand more tools, capable to express more intricate relations. These tools should also be more precise and more formally specified. [Smith 05] undertakes a significant step in this respect: it sets upon the task of improving the reliability and precision of biological and medical ontologies. In short, the diagnostic put by the authors—and with which we happen to concur—is that most existing biological and medical ontologies can be bettered by adopting tools and methods inspired from formal logic and formal ontology. Such an endeavor is seen as bringing about a greater degree of rigor, which fosters interoperability and integration, and ultimately facilitates the handling of biomedical data in an efficient and unambiguous manner by both human operators and especially by automated devices.

The concrete steps undertaken in this respect by [Smith 05] consist in (a) providing a formal (and ontologically sound) support to the relations currently used in most biomedical ontologies—namely *is_a* and *part_of*, and (b) enhancing the list of such relations with new ones in order to compensate for the paucity of expressive means of said ontologies. The former undertakes a reconstruction of these two relations within the language of first-order logic (FOL), considered by the authors to be a sufficiently expressive and rather uncontroversial framework. The latter proceeds by assuming a host of new primitive individual-level relations (i.e. relations between individuals)—e.g. **part_of**, **instance_of**, **located_in**, **has_agent** etc. This background is then used to construct the corresponding class-level relations (i.e. relations between classes), and to define new individual-level and corresponding class-level relations. As a result, a new ontology (RO) emerges. As of January 2007, RO comprises thirteen class-level relations (see http://obofoundry.org/ro/).

While the ACGT MO *uses* RO, its domain-specific requirements call for more *domain-specific relations* which, so far, the RO does not supply, and while we happen to have knowledge of several efforts aimed at enhancing RO, we have nevertheless seen ourselves compelled to take the initiative of defining *new relations*, such that, as of this writing, the number of relations in MO's arsenal amounts to an impressive 242. We will give in the following a few examples of both RO and proper ACGT MO relations, and illustrate their use via ACGT MO's constraints and restrictions.



- acgt:realizes
- acgt:relapseOf
  - acgt:distantRecurrenceOf
  - acgt:regionalRecurrenceOf ↔ acgt:hasRegionalRecurrence
    - acgt:localRecurrenceOf
- acgt:relativeOf ↔ acgt:hasRelative
- acgt:removedBy ↔ acgt:removes
  - acgt:excisedBy
- acgt:removes
  - acgt:excises ↔ acgt:excisedBy
- acgt:revealedBy
- acgt:reveals ↔ acgt:revealedBy
  - acgt:measures
- acgt:sampledBy ↔ acgt:samples
- acgt:sampledFrom
- acgt:samples
- acgt:signedBy
- acgt:situatedIn ↔ acgt:locusOf
- acgt:spatialPositionOf ↔ acgt:hasSpatialPosition
- acgt:subfunctionOf
- acgt:symptomOf ↔ acgt:hasSymptom
- acgt:synchronousWith
- acgt:therapyAimOf
- acgt:toxicityLevelOf
  - acgt:cTCGradeOf ↔ acgt:hasCTCGrade
- acgt:toxicityTypeOf
- acgt:tumorClassOf
  - acgt:mammaryCarcinomaClassOf ↔ acgt:hasMammaryCarcinomaClass
  - acgt:nephroblastomaClassOf ↔ acgt:hasNephroblastomaClass
- ro:adjacentTo
- ro:containedIn ↔ ro:contains
- ro:contains ↔ ro:containedIn
- ro:derivedInto ↔ ro:derivesFrom
- ro:derivesFrom ↔ ro:derivedInto
- ro:hasPart ↔ ro:partOf
- ro:hasParticipant ↔ ro:participatesIn
- ro:locatedIn ↔ ro:locationOf
- ro:locationOf ↔ ro:locatedIn

Figure 8: MO relations sample screenshot

Examples of ACGT MO constraints:

– “Organism *undergoes* some Metabolism” says that for any Organism instance there is a Metabolism instance such that the two instances stand in the **undergoes** relation with each other; here *undergoes* is an MO-specific relation;

- “Organism *hasBirth* exactly 1” says that any Organism instance has exactly one birth; here *hasBirth* is an MO-specific relation;
- “MetabolismByproduct *outcomeOf* some Metabolism” says that for any MetabolismByproduct instance there is a Metabolism instance such that the two instances stand in the **outcomeOf** relation to each other; here *outcomeOf* is an MO-specific relation;
- “OrganismalIndependentContinuant *part of* only Organism” says that an OrganismalIndependentContinuant instance can only be a part of an instance of Organism; here **part of** is an RO relation;
- “Neoplasm *hasMetastasis* min 0” says that a Neoplasm instance can have no less than zero metastases; here *hasMetastasis* is an MO-specific relation.

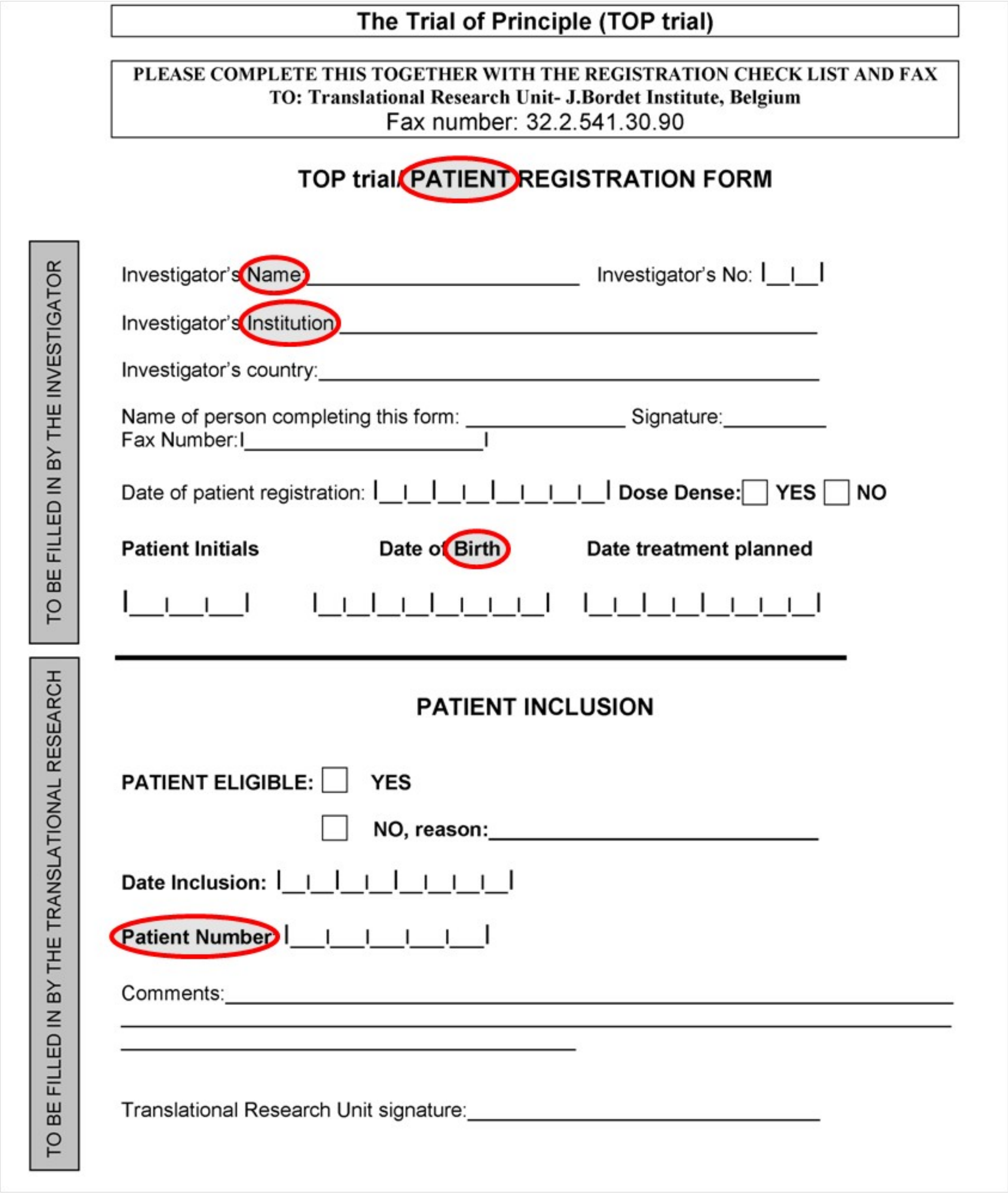
4 Multigranular ACGT MO

Due to the sheer size of its intended domain (clinical trials and biomedicine), the ACGT Master Ontology is a multigranular ontology. This means that it will be representing and cataloguing items and entities at various levels of abstraction (“granularity”), which poses further difficulties from both ontological and logical points of view. One of the main consequences—the only one that will be mentioned here—is that some of the BFO classes at its top will have to be phased out, as they pertain to ontologies that target the same granularity level. These classes are: ProcessAggregate and ObjectAggregate: when taking granularity into account, objects and processes that appeared as units regarded from at a certain level of abstraction may appear as aggregates/conglomerates from the point of view of a lower granularity level and vice versa—hence the ontologically unreliable status of such classes.

5 Sources

The first push toward the creation of the ACGT Master Ontology came via incorporating textbook knowledge into the ontological representation of the domain. It is not therefore surprising that the first sketches of what would become the MO were drawn by Anand Kumar, an ontologist with medical background. A number of textbook resources, as well as existing terminologies, were used during this phase, some of which will be mentioned below, while the rest can be found listed in Deliverable 7.1 (the state of the art review).

The set of Clinical Report Forms (CRFs) provided by our clinical partners has become one of our most important tools. Both the SIOP trial forms and the TOP trial forms were analyzed in great detail—which is not to say that they will not be analyzed and reviewed further—and the relevant classes were extracted and brought to populate the ontology. The whole process was supervised by our clinical partners, whose invaluable feedback came through a great variety of channels, from face to face encounters and discussions, to mailing lists (the ACGT MO boasts its very own dedicated Google group (http://groups.google.com/group/OCInv?pli=1)), Skype conferences etc. Here below (figures 9 and 10) are a few graphic examples of CRF classes that have either found their place in, or are at the root of classes from, the Master Ontology; note that the circled terms have been so highlighted for illustration purposes only, hence in no way should this lead one infer that the non-highlighted terms have been left out.



The Trial of Principle (TOP trial)

PLEASE COMPLETE THIS TOGETHER WITH THE REGISTRATION CHECK LIST AND FAX TO: Translational Research Unit- J.Bordet Institute, Belgium
Fax number: 32.2.541.30.90

TOP trial PATIENT REGISTRATION FORM

TO BE FILLED IN BY THE INVESTIGATOR

Investigator's Name____________________ Investigator's No: |__|__|

Investigator's Institution__________________________________

Investigator's country:____________________________________

Name of person completing this form: _____________ Signature:_________
Fax Number:|____________________|

Date of patient registration: |__|__|__|__|__|__|__|__| **Dose Dense:** ☐ **YES** ☐ **NO**

Patient Initials **Date of Birth** **Date treatment planned**

|__|__|__| |__|__|__|__|__|__|__|__| |__|__|__|__|__|__|__|__|

PATIENT INCLUSION

TO BE FILLED IN BY THE TRANSLATIONAL RESEARCH

PATIENT ELIGIBLE: ☐ **YES**

☐ **NO, reason:**__________________________

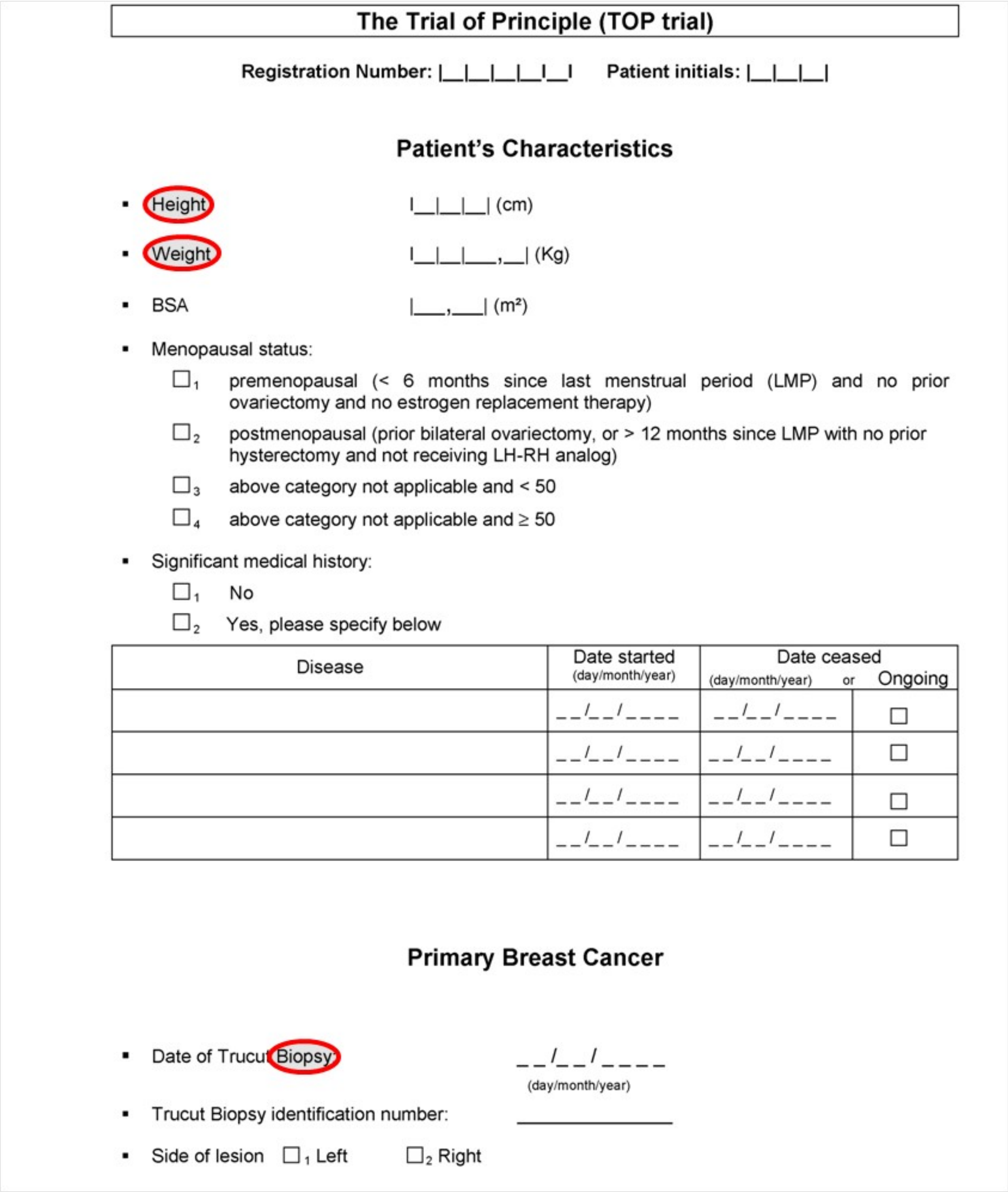
Date Inclusion: |__|__|__|__|__|__|__|__|

Patient Number |__|__|__|__|__|

Comments:____

Translational Research Unit signature:____________________________

Figure 9: Classes from the TOP trial CRFs



The Trial of Principle (TOP trial)

Registration Number: |__|__|__|__|__| **Patient initials: |__|__|__|**

Patient's Characteristics

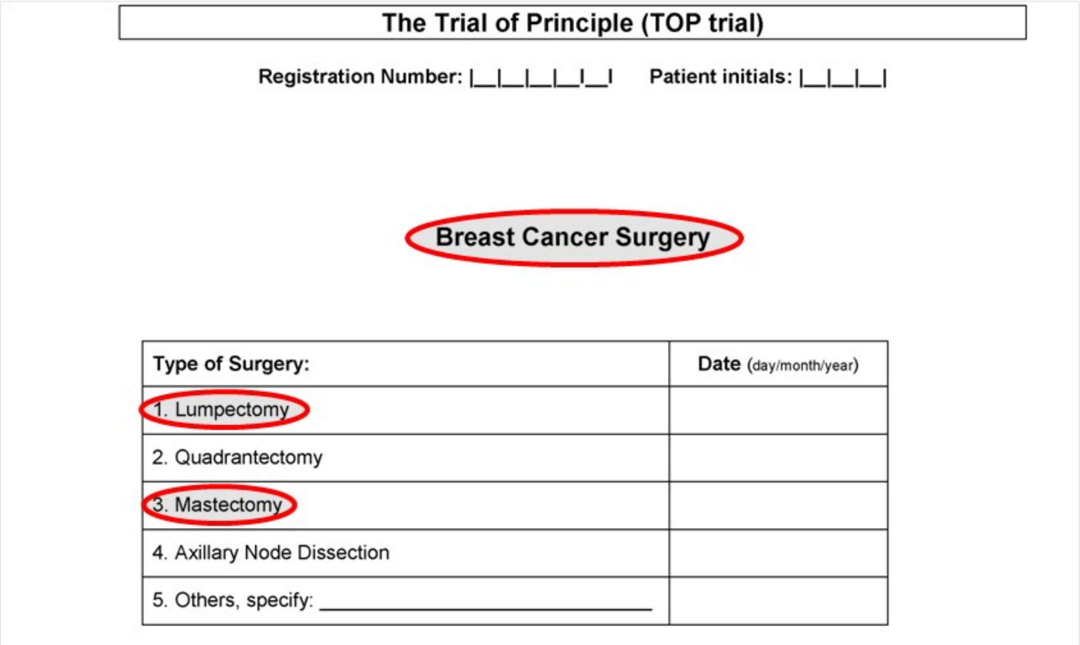
- Height |__|__|__| (cm)
- Weight |__|__|___,__| (Kg)
- BSA |___,___| (m²)
- Menopausal status:
  - ☐$_1$ premenopausal (< 6 months since last menstrual period (LMP) and no prior ovariectomy and no estrogen replacement therapy)
  - ☐$_2$ postmenopausal (prior bilateral ovariectomy, or > 12 months since LMP with no prior hysterectomy and not receiving LH-RH analog)
  - ☐$_3$ above category not applicable and < 50
  - ☐$_4$ above category not applicable and ≥ 50
- Significant medical history:
  - ☐$_1$ No
  - ☐$_2$ Yes, please specify below

Disease	Date started (day/month/year)	Date ceased (day/month/year) or	Ongoing
	__/__/____	__/__/____	☐
	__/__/____	__/__/____	☐
	__/__/____	__/__/____	☐
	__/__/____	__/__/____	☐

Primary Breast Cancer

- Date of Trucut Biopsy: __/__/____ (day/month/year)
- Trucut Biopsy identification number: ____________
- Side of lesion ☐$_1$ Left ☐$_2$ Right

Figure 10: More classes from the TOP trial CRFs



The Trial of Principle (TOP trial)

Registration Number: |__|__|__|__|__| **Patient initials:** |__|__|__|

Breast Cancer Surgery

Type of Surgery:	**Date** (day/month/year)
[✓ 1. Lumpectomy]	
2. Quadrantectomy	
[✓ 3. Mastectomy]	
4. Axillary Node Dissection	
5. Others, specify: ____________________________	

Figure 11: More classes from the TOP trial CRFs

Among other resources used to populate the ACGT ontology we can mention several widely used cutting edge cancer and pathology books such as [DeVita 01] (www.LWWoncology.com), [Schulz 05] and [Kumar 04]. Many existing specialized ontologies and controlled vocabularies have also decisively influenced the MO taxonomy: the TNM cancer staging system (see figure 12), the Foundational Model of Anatomy (FMA (http://sig.biostr.washington.edu/projects/fm/index.html)) (see figure 13), the Gene Ontology (GO (http://www.geneontology.org/)). Last but not least, a great deal of insight has been drawn from other contemporary efforts targeted at developing clinical trial ontologies such as the OBI group (Ontology of Biomedical Investigation (http://obi-ontology.org/page/Main_Page)), the OCI group (Ontology of Clinical Investigation) and many other components of the OBO Foundry (http://www.obofoundry.org/). At this point we should again emphasize that all this effort aimed at amassing and compiling a significant portion of medical knowledge under the guise of an ontology has been undertaken under the careful and constant supervision of our specialist medical partners.

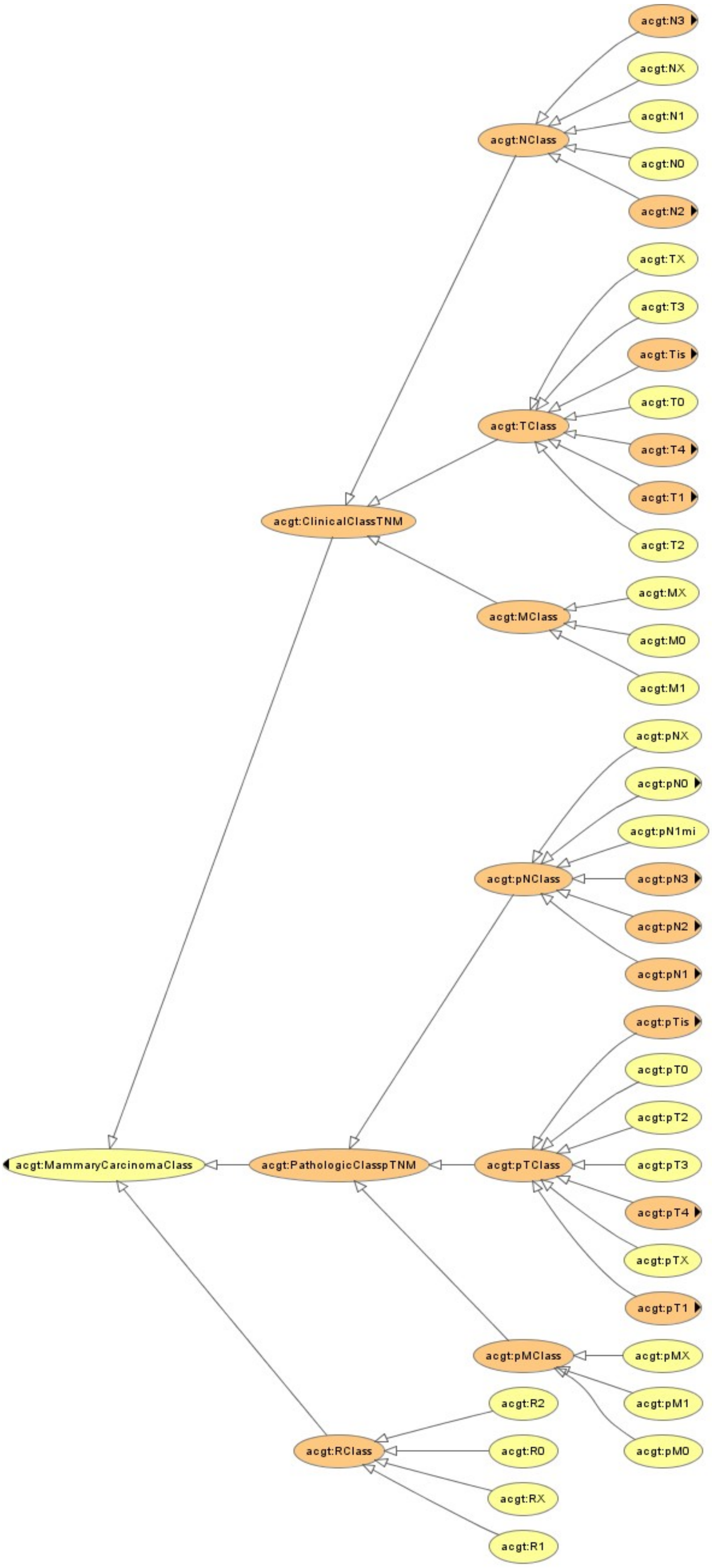


Figure 12: The TNM cancer staging as reflected in the MO

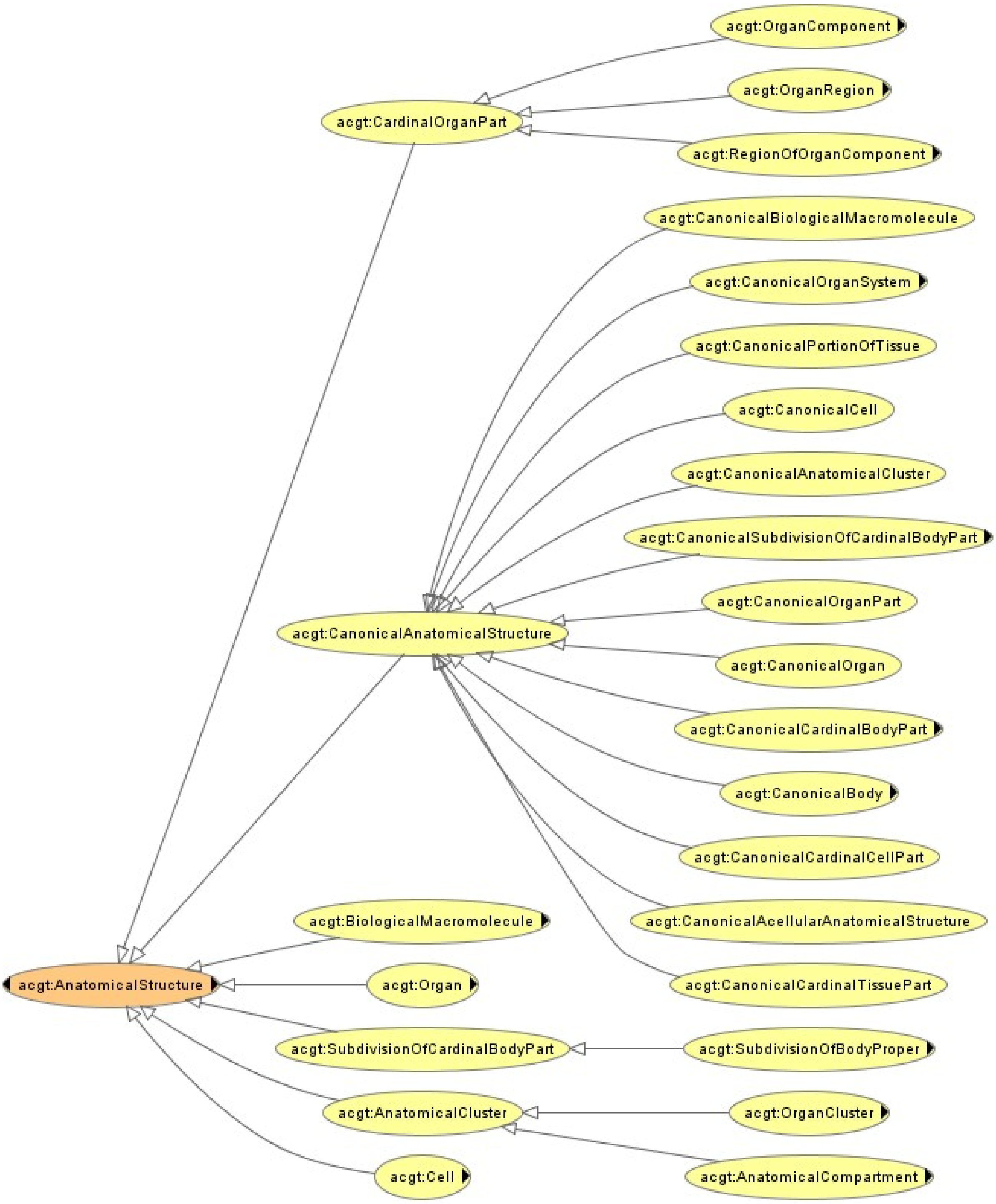


Figure 13: An FMA sample as reflected in the MO

6 Methodology

Given the extent of the project at hand, selecting and sticking with a *single* ontology development method of the many available has been a challenging task. For the most part we have rather loosely followed the development scheme described in [Fernandez-Lopez 97], and dubbed “Methontology.” The following is a cursory presentation of Methontology adapted from [Gomez-Perez 04]; for details, one should refer to either this, or the original article ([Fernandez-Lopez 97]).

Methontology recognizes three groups of activities (figure 14), which intertwine rather arbitrarily during the ontology development life cycle (figure 15):

1. *Project Management Activities* include planning, control and quality assurance. *Planning/scheduling* identifies what tasks are to be performed, how they will be arranged, how much time, and what other resources are needed for their completion. This activity is indispensable for ontologies that need to (re)use already existing ontologies, or ontologies that require levels of abstraction and generality. *Control*, guarantees that planned tasks are completed in the manner in which they were intended to be performed. *Quality Assurance*, ensures that the quality of each and every product outputted (ontology, software and documentation) is satisfactory.
2. *Development-oriented Activities* comprise specification, conceptualization, formalization, implementation and maintenance. *Specification* states why the ontology is being built, what its intended uses are, and who are the targeted end-users. *Conceptualization* structures the domain knowledge as meaningful models at the knowledge level. *Formalization* transforms the conceptual model into a formal or semi-computable model. *Implementation* builds computable models in a computational language. Finally, *maintenance* updates and corrects the ontology.
3. *Ontology Support Activities* consist of a series of actions performed at the same time as development-oriented activities, without which the ontology could not be built. This step incorporates *knowledge acquisition*, *evaluation*, *integration*, *documentation* and *configuration management*. *Knowledge Acquisition* aims, obviously, at acquiring knowledge of the targeted domain. *Evaluation* makes a technical judgment of the ontologies, their associated software environments and documentation with respect to a frame of reference during each phase and between phases of their life cycle. *Integration* of ontologies is required when building a new ontology that reuses other already available ontologies. *Documentation* seeks to detail each and every one of the phases completed and products generated. *Configuration Management* records all the versions of the documentation, software and ontology code to control the changes.

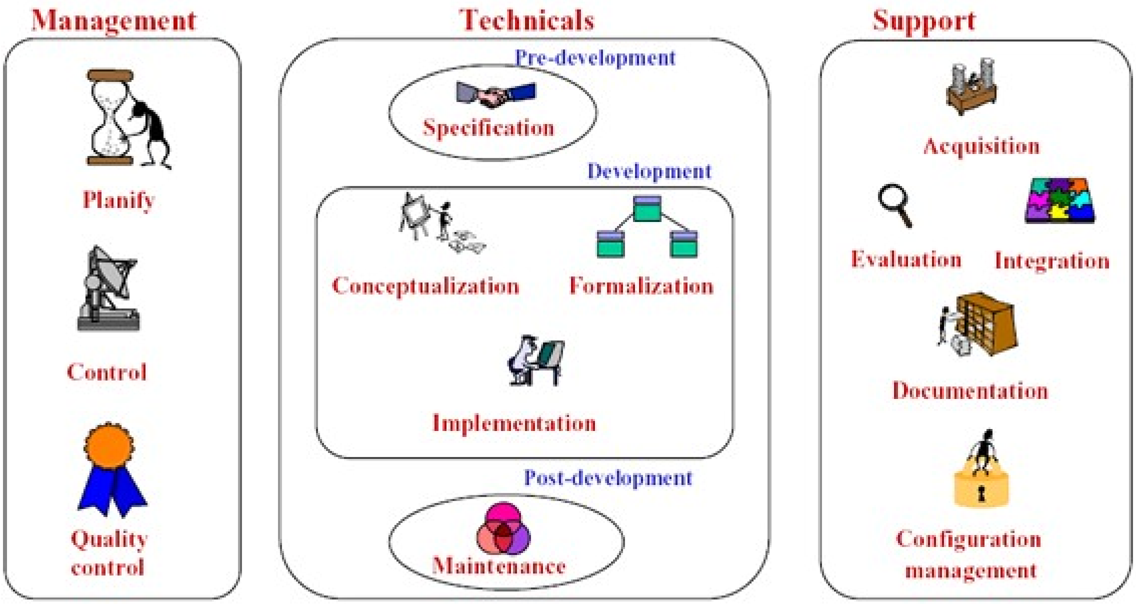


Figure 14: Ontology Development Process [Corcho 05]

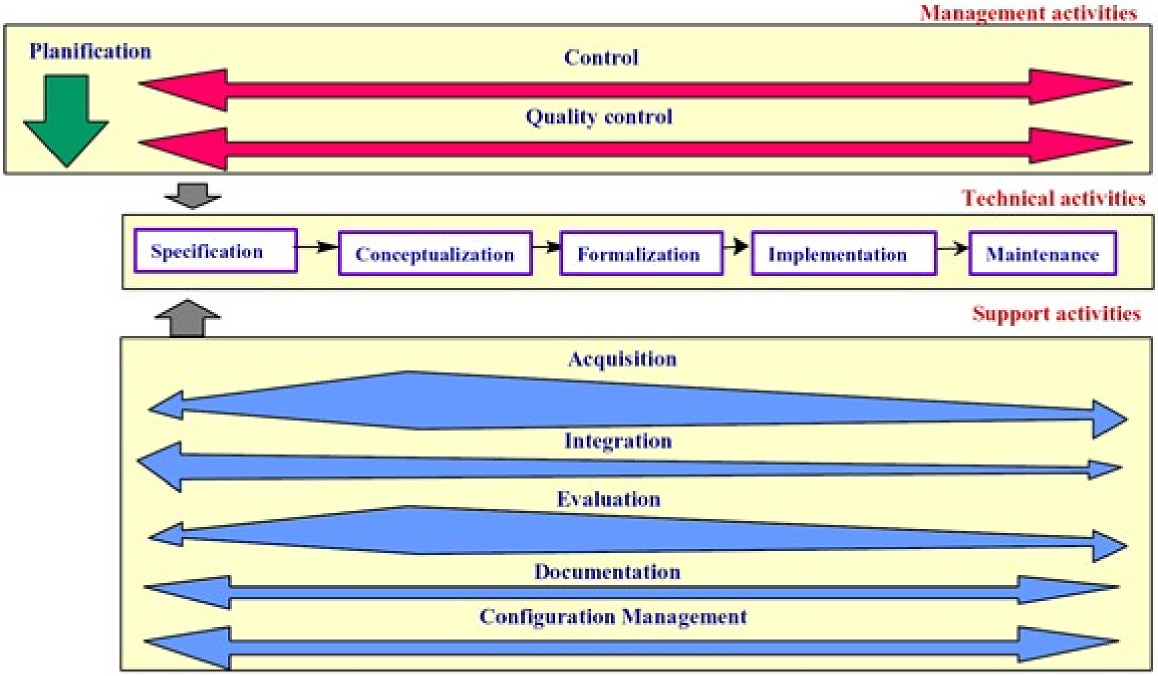


Figure 15: Methontology ontology development process life cycle [Corcho 05]

During the development cycle of the ACGT MO some of these steps as have been documented and published elsewhere (see previous IFOMIS deliverables, conference proceedings, and journal articles). Obviously the most important phase is the actual development phase (6 above), and, within this, the conceptualization step stands out as crucially important. We will present it in some detail in the following, as it has been outlined in [Gomez-Perez 04], and to the extent to which the steps involved are relevant for our case.

The objective of the conceptualization phase is to organize and structure the knowledge acquired during the knowledge acquisition stage, using external representations that are independent of the knowledge representation and implementation paradigms in which the ontology will be further formalized and implemented. An informally perceived view of a domain is thus converted into a semi-formal model using *intermediate representations* (IRs) based on tabular and graph notations. These intermediate representations (class, attribute, relation, axiom and rule) are essential components, as they can be understood by *both* domain experts *and* ontology developers. They, therefore, bridge the gap between people's domain perception and ontology implementation languages.

In order to build a consistent and complete conceptual model, the conceptualization activity defines a set of tasks that should be executed in succession. These tasks increase, step by step, the complexity of the intermediate representations used to build the conceptual model. This way it is easier to ensure a consistent and complete conceptual model:

1. One starts by putting together a *glossary of terms* to be included in the ontology, as well as their natural language definitions and their synonyms and acronyms. Terms are identified following a middle-out strategy. The core terms are identified first and then they are specialized and generalized as required. This strategy provides a balanced set of terms because detail only arises as necessary and higher level categories are built naturally.
2. Then, the terms are classified into one or more *taxonomies* of concepts, where a class is an abstraction for one or more terms. The “subclass_of” taxonomic relation is used.
3. *Binary relations* are used to define the ad hoc relations between classes of the ontology and also with classes of other ontologies. Relations are determined by their name and the source and target classes.
4. The *class dictionary* is built. It describes each class by stating the relations that have it as their domain.
5. The class dictionary is detailed. For each relation, one specifies the cardinality, inverse relation and mathematical properties (symmetric, transitive, functional etc.). The outcome is the binary relation table.
6. Once concepts, taxonomies and relations have been defined, formal axioms and rules are used for constraint checking. *Axioms* are logical expressions that are always true and are normally used to specify constraints. They are defined informally in textual form and formally in first order logic. Moreover, all the classes and relations used in the definitions are highlighted. *Rules* are generally used to infer knowledge in the ontology, such as relation instances. Rules are also defined informally and formally and the related classes and relations are highlighted.

7 Application: SPARQL Queries

Among the most immediate applications of the MO we will present in the following some aspects pertaining to the mapping process; more specifically, we will give the MO translation of three queries destined to query TOP data (http://wiki.healthgrid.org/ACGT:TOP_CRF_database). These translations have been worked out with the assistance of ACGT partners directly involved in the mapping process, and are the result of numerous iteration steps and heated debates.

Microarray Identifiers SPARQL query:

- Natural language formulation: “Retrieve the identifiers of the microarray files for each patient”
- SPARQL query in MO terms:

```
SELECT ?patient ?arrayID {
   ?patient a acgt:HumanBeing ;
            acgt:undergoes ?biopsy .
   ?biopsy a acgt:Biopsy ;
           acgt:hasOutcome ?tissueSample .
   ?tissueSample a acgt:HistologicalSample ;
                 acgt:partOf ?microarray .
   ?microarray a acgt:TissueMicroarray ;
               acgt:hasIdentifier ?identifier ;
   ?identifier a acgt:Identifier ;
               acgt:hasStringValue ?arrayID .
 }
```

- Diagram:

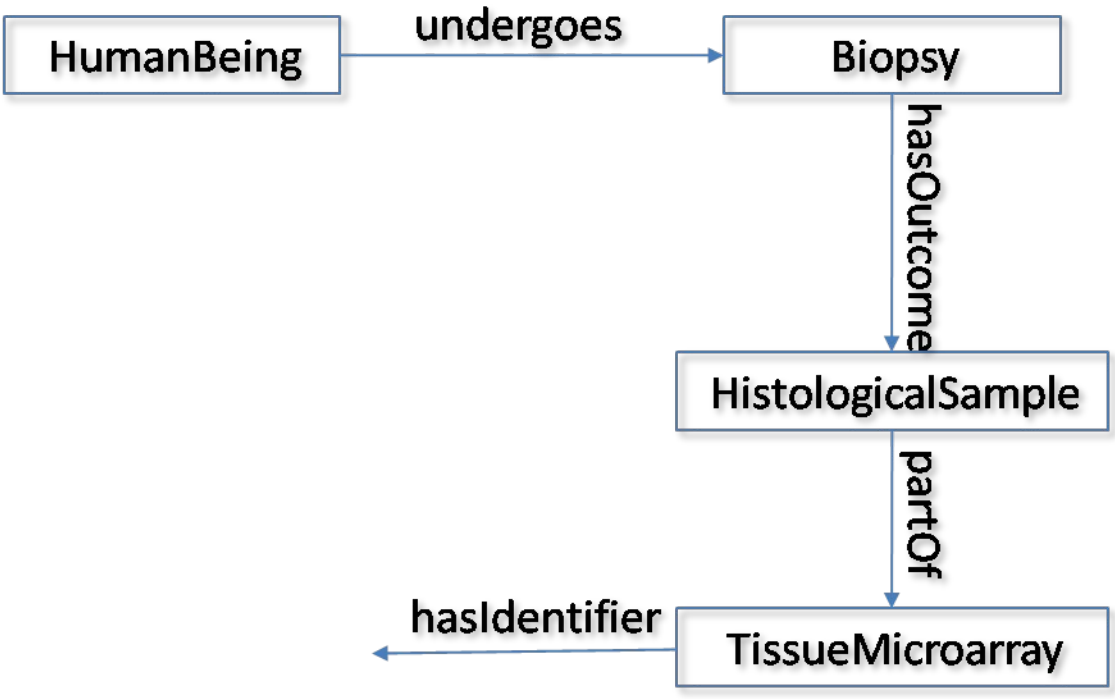


Figure 16: Microarray query

Patient Age at Registration SPARQL query:

- Natural language formulation: “Determine patients' age from the registration dates and the patients' birth date”

- SPARQL query in MO terms:

```
SELECT ?patient ?birthDate ?regDate {
   ?patient a acgt:HumanBeing ;
            acgt:participatesIn ?trial ;
            acgt:hasBirthDate ?birthDate .
   ?trial a acgt:ClinicalTrial ;
          acgt:hasIdentifier "TOP" .
   ?registration a acgt:Registration ;
                 acgt:hasProcessEnd ?endOfReg ;
                 acgt:hasPatient ?patient ;
                 acgt:administrativeProcessOf ?trial .
   ?endOfReg acgt:hasDate ?regDate .
 }
```

- Diagram:

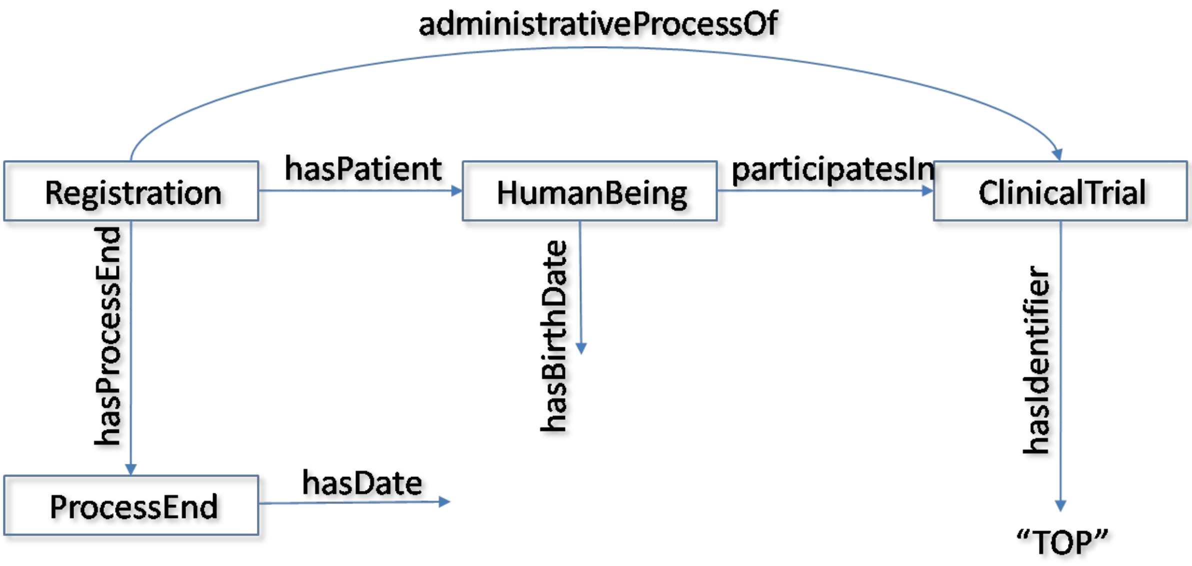


Figure 17: Patient age query

Residual Tumor Size and Pathological Complete Response SPARQL query:

- Natural language formulation: "Retrieve the following information for the patients: (a) residual tumor size after neo-adjuvant treatment and (b) pathological complete response (PCR) status"
- SPARQL query in MO terms:

```
SELECT ?patient ?tumorWidth ?tumorHeight ?pcrYN {
   ?patient a acgt:HumanBeing ;
            acgt:undergoes ?chemotherapy .
   ?neoplasm a acgt:Neoplasm ;
             acgt:undergoes ?tumorDiagnosis ;
```

```
            acgt:partOf ?patient ;
            acgt:undergoes ?chemotherapy .
   ?chemotherapy a acgt:ChemoTherapy ;
                 acgt:precedes ?tumorDiagnosis .
   ?tumorDiagnosis a acgt:DiagnosticProcess ;
                   acgt:reveals ?widthQuality ;
                   acgt:reveals ?heightQuality ;
                   acgt:reveals ?pcrStatus .
   ?widthQuality a acgt:Width ;
                 acgt:hasFloatValue ?tumorWidth .
   ?heigthQuality a acgt:Height ;
                 acgt:hasFloatValue ?tumorHeight .
   ?pcrStatus a acgt:pCRStatus ;
              acgt:hasBoleanValue ?pcrYN .
 }
```

- Diagram:

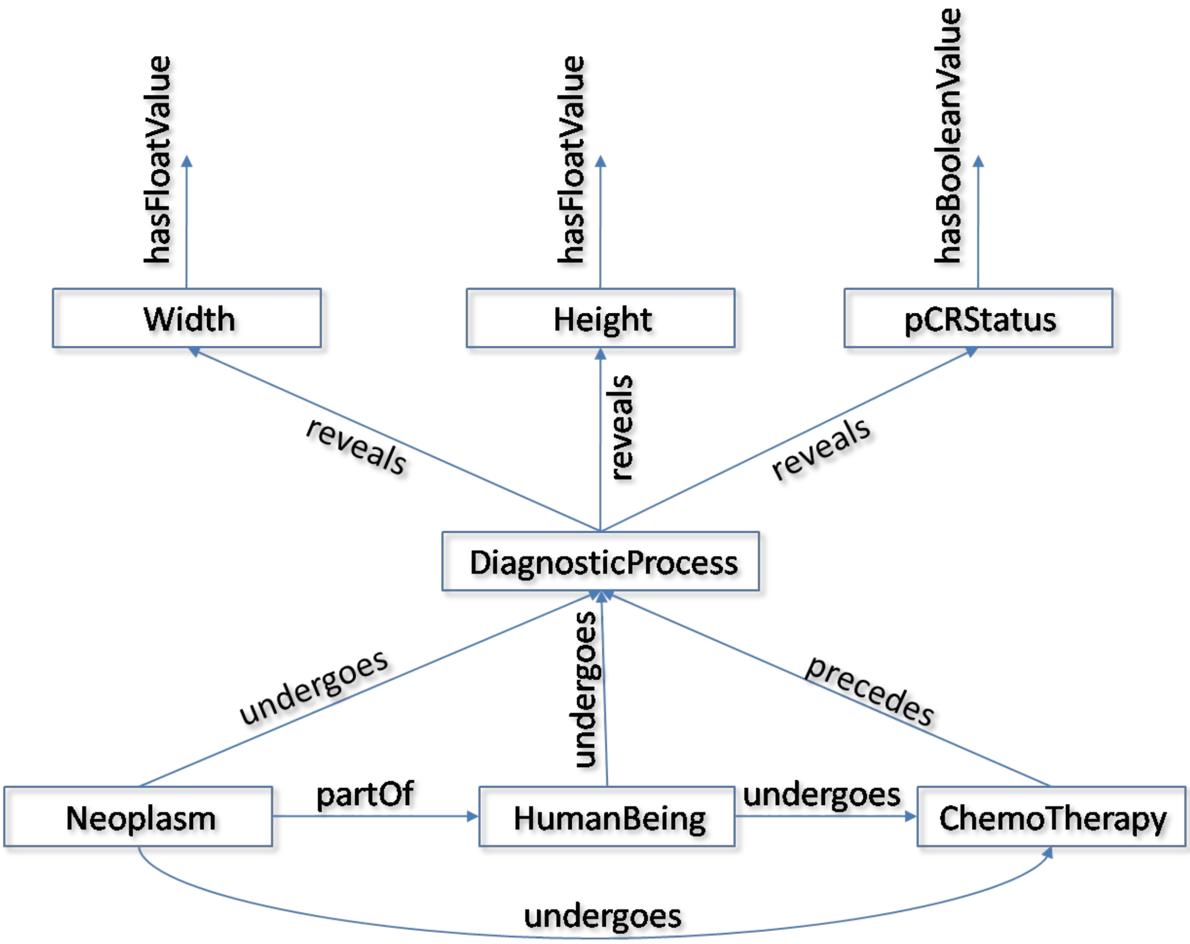


Figure 18: PCR query

8 Bibliography

[DeVita 01] DeVita, V. et. Al., Cancer: Principles and Practice of Oncology, 6th edition. Lippincott Williams & Wilkins, 2001

[Schulz 05] Schulz, W., Molecular Biology of Human Cancers: An Advanced Student's Textbook. Springer, 2005

[Kumar 04] Kumar, V., Robbins Basic Pathology, 7th edition. Saunders, 2004

[Gruber 93] T.R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5:199–220, 1993

[Rector and Welty 05] A. Rector and C. Welty. Simple part–whole relations in OWL ontologies. Technical report, World-Wide Web Consortium (W3C), Working Group Note, http:// www.w3.org/2001/sw/BestPractices/OEP/SimplePartWhole/, 2005

[Smith 05] Smith, B. et. al., "Relations in biomedical ontologies," *Genome Biology* 2005, **6:**R46

[Smith 06] Smith, B. et. al., "Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain," *Proceedings of KR-MED 2006*

[Smith 07] Smith, B. et. Al., "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nature Biotechnology* 25, 1251 - 1255

[Gómez-Pérez 04] Gómez-Pérez, A., Fernández-López, M., Corcho, O. "Ontological Engineering." Springer, London, 2004

[Lassila 01] Lassila, O., McGuinness, D. "The role of frame-based representation on the Semantic Web." Technical Report KSL-01-02, Knowledge System Laborartory, Stanford University, Stanford, 2001

[Wittekind 05] Wittekind, C. et. al., TNM Klassifikation maligner Tumoren. Springer, 2005

[Fernandez-Lopez 97] Fernandez-Lopez, M.; Gomez-Perez, A. & Juristo, N.: "Methontology: from ontological art towards ontological engineering". "Proc. Symposium on Ontological Engineering of AAAI," 1997

[Corcho 05] Corcho, O.; Fernandez, M.; Gomez-Perez, A. & Lopez-Cima, A.: "Building Legal Ontologies with METHONTOLOGY and WebODE". In Benjamins, R.; Casanovas, P.; Breuker, J. & Gangemi, A. (ed.): "Law and the Semantic Web". *Springer-Verlag,* LNAI No. 3369, pp. 142-157, 2005

[Spear 06] Spear, A., Ontology for the Twenty First Century: An Introduction with Recommendations, BFO manual, 2006

Appendix 1 - Abbreviations and acronyms

BFO	Basic Formal Ontology
BIRNLex	Biomedical Informatics Research Network Lexicon
CIDOC	Committee on Documentation of the International Council of Museums
CRF	Case Report Form

CRM	Conceptual Reference Model
DOLCE	Descriptive Ontology for Linguistic and Cognitive Engineering
FMA	Foundational Model of Anatomy
FOL	First Order Logic
GO	Gene Ontology
HL7	Health Level Seven
MeSH	Medical Subject Headings
MO	Master Ontology
NCIT	National Cancer Institute Thesaurus
OBI	Ontology of Biomedical Investigation
OBO	Open Biomedical Ontologies
OCI	Ontology of Clinical Investigation
OWL	Web Ontology Language
RO	Relation Ontology
SIOP	International Society of Pediatric Oncology
SPARQL	SPARQL Protocol and RDF Query Language (recursive acronym)
SUMO	Suggested Upper Merged Ontology
TNM	TNM Classification of Malignant Tumors
TOP	Trial of Principle
UMLS	Unified Medical Language System

Appendix 2 - Glossary

Adequatism This is the position that a good theory of reality must do justice to all of the different phenomena that reality contains. In opposition to the tendency to attempt to reductively explain higher level macroscopic phenomena in terms of “more basic” or fundamental components of reality, adequatism entails that the entities in any given domain of reality be taken seriously on their own terms first. Thus, just as an ontology of physics should be about atoms and sub-atomic particles, and an ontology of chemical reactions should include the existence of various kinds of elements and compounds, so an ontology of biological phenomena should include the existence of, at various levels, cells, organs, biological systems and organisms, as well as populations and environments. The goal of adequatism is to do justice to the vast array of different kinds of entities that exist in the world, in different domains and at different levels of granularity, rather than ignoring them or attempting to explain them away.

Fallibilism Fallibilism involves commitment to the idea that, although our current scientific theories are the best candidates we have for the truth about reality, it may nevertheless be the case that portions of our current knowledge are incorrect, hence our current purported reality representations are *not* representations after all. The fallibilist maintains that it is a matter of empirical investigation what the facts of reality are, and recognizes that empirical investigation is an ongoing, open-ended, experimental process.

Perspectivalism Perspectivalism involves the recognition that reality is a complex and variegated phenomenon. While not all purported representations of reality are good, because some are accurate to the facts of reality and some are not, there are nevertheless many different equally good representations (good in the sense of being true), precisely in that they capture different and important features of one and the same reality, that is, they capture competing angles of investigation.

Property An OWL property is a binary relation. Two types of properties are distinguished: first, datatype properties, relations between instances of classes and RDF literals and XML Schema datatypes. Secondly, object properties, relations between instances of two classes [22].

Property Restrictions When properties are defined there are a number of ways to restrict those relation. The domain (i.e. subject) and range (i.e. object) of the properties can be specified. The property can be defined to be a specialization (subproperty) of an existing property and also more elaborate restrictions (e.g. cardinality restrictions) are possible [18].

Realism ‘Realism’ can be defined as the view according to which reality and its constituents exist independently of our (linguistic, conceptual, theoretical, cultural) representations. Realism is the thesis that the things that scientific knowledge is about are in fact real, mind-independent things. Thus, ontologies are representations of reality, not representations of people’s concepts or mental representations of reality.