A collaborative methodology for developing a semantic model for interlinking Cancer Chemoprevention linked-data sources

Editor(s): Mikel Egaña Aranguren, Technical University of Madrid, Spain; Michel Dumontier, Carleton University, Canada; Jesualdo Tomás Fernández Breis, University of Murcia, Spain.
Solicited review(s): Mari Carmen Suárez-Figueroa, Technical University of Madrid, Spain; Alejandro Rodríguez González, Polytechnic University of Madrid, Spain; Erick Antezana, Norwegian University of Science and Technology, Norway; Robert Stevens, University of Manchester, UK; Iker Huerga, Linkatu, Spain.

Dimitris Zeginis[a,b], Ali Hasnain[c], Nikolaos Loutas[a,b,c], Helena Futscher Deus[c], Ronan Fox[c], Konstantinos Tarabanis[a,b]
[a] *Centre for Research and Technology Hellas, Thessaloniki, Greece*
[b] *Information Systems Lab, University of Macedonia, Thessaloniki, Greece*
{zeginis, nlout, kat}@uom.gr
[c] *National University of Ireland, Galway, Digital Enterprise Research Institute, Galway, Ireland*
firstname.lastname@deri.org

Abstract. This paper proposes a collaborative methodology for developing semantic data models. The proposed methodology for the semantic model development follows a "meet-in-the-middle" approach. On the one hand, the concepts emerged in a bottom-up fashion from analyzing the domain and interviewing the domain experts regarding their data needs. On the other hand, it followed a top-down approach whereby existing ontologies, vocabularies and data models were analyzed and integrated with the model. The identified elements were then fed to a multiphase abstraction exercise in order to get the concepts of the model. The derived model is also evaluated and validated by domain experts. The methodology is applied on the creation of the Cancer Chemoprevention semantic model that formally defines the fundamental entities used for annotating and describing inter-connected cancer chemoprevention related data and knowledge resources on the Web. This model is meant to offer a single point of reference for biomedical researchers to search, retrieve and annotate linked cancer chemoprevention related data and web resources. The model covers four areas related to Cancer Chemoprevention: i) concepts from the literature that refer to cancer chemoprevention, ii) facts and resources relevant for cancer prevention, iii) collections of experimental data, procedures and protocols and iv) concepts to facilitate the representation of results related to virtual screening of chemopreventive agents.

Keywords: Collaborative model development; Common data model, Cancer Chemoprevention; Linked Data; HCLS

1. Introduction and motivation

In all scientific areas there exists an increasing amount of information available to assimilate. In some fields, such as biology, this increase is even more obvious because of the high-throughput lab techniques and electronic publishing technologies used. The result is that science increasingly depends on computers to store, access, integrate, and analyze data. In order to exploit the power of semantic web and linked-data technologies the knowledge has to be formalized. The first step in formalizing knowledge is to define an explicit data model.

In ontology engineering literature there exist many methodologies for creating ontologies and semantic data models (*e.g.* [1], [2]), which are mainly based on competency questions to determine the domain and scope of the ontology. This approach lacks interaction during ontology development with the end-users of the ontology so increasing the risk of coming up with a well-formed ontology that may be not practically usable by the end-users.

In order to face this challenge, input and feedback of the end-users of the ontology is required, not just after the ontology creation but also at all the intermediate steps followed from the specification of the ontology scope to the final result. For this reason there is a need for a new collaborative methodology for ontology and semantic data model development that heavily relies on user feedback, not only at the specification of the semantic model scope, but at all the methodology steps.

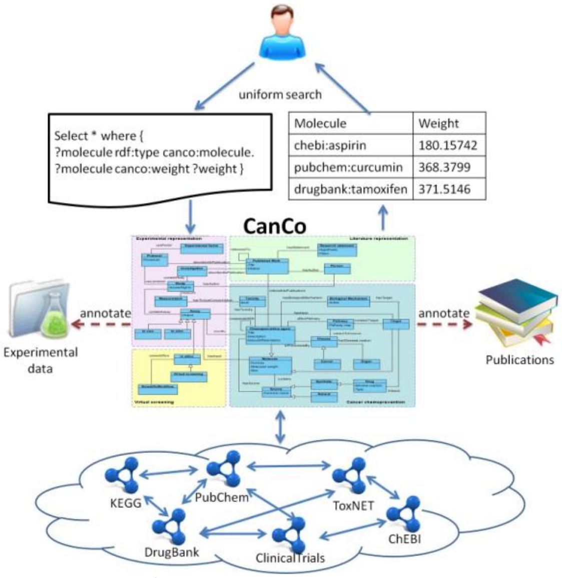
This paper proposes such a collaborative methodology for developing semantic data models. In order to clarify the proposed methodology we employed it for building a cancer chemoprevention semantic model (CanCo). Cancer chemoprevention is defined as the use of natural, synthetic, or biologic chemical agents to reverse, suppress, or prevent the carcinogenic progression to invasive cancer [3, 4]. It is considered as one of the most promising areas in current cancer research [5].

The challenges encountered were related to determining the scope of the semantic model and creating a model practically usable by the biomedical researchers.

As part of this work, we have analyzed approximately 70 biomedical data sources (vocabularies, ontologies, linked datasets and reference data) found in the literature but they are generic enough and do not fully cover the peculiarities of the cancer chemoprevention domain. For example, the Experimental Factor Ontology (EFO) [6] and the Ontology for Biomedical Investigations (OBI) [7] cover aspects related to the biomedical experiments, but they do not connect the experiments to cancer chemoprevention processes. Moreover, the Gene Ontology (GO) [8] and BioPax [9] aim at standardizing the representation of genes and pathways respectively, but they do not relate them with the action of a chemopreventive agent. Therefore, there is lack of an ontological model clearly designed for specifically targeting the Cancer Chemoprevention domain.

Data relevant to cancer chemoprevention is typically spread across a very large number of heterogeneous data sources, including ontologies, knowledge bases, linked datasets, databases with experimental results and publications. The Cancer Chemoprevention semantic model unifies all these data and works as a “glue” between them allowing the querying of data across sources with a single search (by linking the existing Life Sciences LOD Cloud) and the annotation of data (experimental data and publications) related to cancer chemoprevention (Fig. 1).

The remainder of this paper is organized as follows. Section 2 presents the related work on methodologies for ontology development. Section 3 introduces the collaborative methodology for developing semantic models. Section 4 presents the CanCo showcase that demonstrates the proposed methodology. Finally, in Section 0 we conclude the paper and discuss future research directions.



```
Select * where {
?molecule rdf:type canco:molecule.
?molecule canco:weight ?weight }
```

Molecule	Weight
chebi:aspirin	180.15742
pubchem:curcumin	368.3799
drugbank:tamoxifen	371.5146

Fig. 1 The role of the Cancer Chemoprevention semantic model

2. Related work

This section reviews existing methodologies for ontology development. Grüninger and Fox [10] proposed an ontology design and evaluation methodology while developing the TOVE (Toronto Virtual Enterprise) project ontology. They use motivating scenarios and a set of natural language questions that the ontology needs to be able to answer. These questions are called competency questions and are used to determine the scope of the ontology, to extract the main concepts of the ontology and to evaluate the ontology.

Uschold and King [11] propose a methodology for development of ontologies that comprise of four phases. The first step is the definition of the purpose and scope of the ontology, the second steps is the conceptualisation/building/integration of the ontology. For the conceptualisation they use the middle-out approach since it allows the identification of the primary concepts of the ontology. The thirds step is the evaluation of the developed ontology by comparing it to its purpose and scope. The final step is the documentation of the ontology's classes and properties. This methodology defines well grounded steps for an ontology creation but lacks interaction with the end users of the ontology.

Öhgren and Sandkuhl [12] define a similar methodology focusing on the reuse of fragments of existing ontologies and at instruction-like definition of all steps of the development process. In this methodology there is also no active involvement of the ontology's end-users.

METHONTOLOGY [1] consists of six steps: *i)* the ontology specification that determines the scope and granularity of the ontology, *ii)* the knowledge acquisition, *iii)* the conceptualization where the concepts of the ontology are collected, *iv)* the integration of existing ontologies, *v)* the implementation of the ontology and *vi)* the evaluation. In METHONTOLOGY the end-users are involved at the knowledge acquisition phase (e.g. brainstorming, interviews) but they are not actively involved throughout the development process by contributing or providing feedback.

Li et al. [13] propose a similar in spirit methodology that is assisted by a semi-automatic acquisition tool.

A scenario-based methodology is proposed by NeOn [14, 15]. This methodology supports the collaborative aspects of ontology development and reuse, as well as the dynamic evolution of ontology networks in distributed environments. In these scenarios end-users are actively involved.

On-To-Knowledge [16] focuses on a process oriented methodology for developing ontology-based knowledge management systems. It introduces two orthogonal processes with feedback loops, the Knowledge Processes that circles around the usage of ontologies and the Knowledge Meta Processes that guides their initial set up.

A methodologies that support collaborative ontology development is DILIGENT [17]. It proposes a distributed setting to engineer and create ontologies by domain experts with the help of a fine-grained methodological approach based on Rhetorical Structure Theory.

Finally, Villazón-Terrazas et. al [18] proposes a set of methodological guidelines on how to publish data as Linked Data. These guidelines focus more on the reuse (complete or partial) of existing linked-data sources, and no focus is given to the collaborative design of an ontology.

In summary, most of the current ontology development methodologies adopt a workflow of specification, conceptualization, implementation and evaluation but they lack collaboration and actively involvement of the end-users. Only NeOn and DILIGENT involve the end-users, but they do not focus on the collaborative aspect of ontology development. DILIGENT focuses on the distributed setting of the development while NeOn focuses on scenarios for ontology development and provides a planning for each particular case (this plan includes concrete steps to be followed).

3. A collaborative methodology for semantic model development

The methodology proposed by this paper builds upon the methodologies and guidelines discussed in the related work. The focus of the methodology is on the collaborative development of an ontology by defining concrete steps to be followed. The novel part of the approach is the active engagement of the domain experts during the actual development of the model (specification and conceptualization) and not just their limited involvement in the model evaluation.

The methodology is more "dedicated" to Health Care Life Sciences ontologies and adopts a "meet-in-the-middle" approach where concepts emerged both in a bottom-up (i.e. analyzing the domain and interviewing the domain experts regarding their data needs) and top-down (i.e. analyze and integrate existing ontologies, vocabularies and data models) fashion. Specifically, it comprises of the following phases: specification, top-down and bottom-up conceptualization, implementation, and evaluation (Fig. 2). The phases are discussed in detail below.

The **specification** of the semantic model is the first phase to be carried out. It is very critical to carefully design the specification since at this step the scope and the requirements of the semantic model are defined. Both the domain experts and the ontology engineers are involved in this phase. This facilitates the identification of the knowledge that should be represented in the semantic model, and the evaluation of the model by detecting the satisfied requirements. The outputs of the specification phase are:

- The **scope** of the semantic model that defines the main function that the semantic model should have and the domain it is intended to cover. (*e.g.* cancer chemoprevention).

- The **end-users** of the semantic model, namely the actual beneficiaries of the model (*e.g.* biomedical researchers).
- The **end-uses** define the way that the semantic model will be used (*e.g.* searching in multiple resources).
- The **model requirements** are formed in terms of competency questions (*e.g.* what is the weight of a molecule?) that have to be answered by the model.

The domain experts (who in our case also overlap with the end-users) are actively involved in all aforementioned steps. They contribute their input and feedback through brainstorming, interviews and completing questionnaires. The most effective and efficient method to contact the domain experts depends on their knowledge, demographic characteristics etc. Usually a combination of these methods produces the desired result.

The core phase during the development of the semantic model is the **conceptualization**. In this phase the concepts and relationships of the model are identified. We propose a “meet-in-the-middle” approach. On the one hand, relevant concepts emerge in a bottom-up fashion by analyzing the domain and the model specification. On the other hand, a top-down approach is followed through analysis of relevant existing ontologies and data models. The conceptualization steps are the following:

- Identification of the **core concepts** that come out from the **specification** phase by analyzing the scope and the requirements identified. Specifically, the core concepts are identified by manually analyzing the competence questions defined at the specification. For example based on the competence question “what is the weight of a molecule?” the concept “molecule” is identified. These concepts act as a “seed” for the semantic model (*bottom-up*).
- Identification of **related models and ontologies**, **analyze** them and **reuse** concepts. An important part of this step is the identification of related models and ontologies that can be reused. To do so, search engines and repositories specific to the target domain should be investigated. The related models and ontologies are identified with the help of the domain experts. Once related ontologies are found, they are analyzed collaboratively with the domain experts to identify the concepts that can be reused. In order to determine the relevance of a model a set of criteria is used [14]: *i*) scope of the ontology, *ii*) purpose of the ontology, *iii)* functional and non-functional requirements covered. (*top-down*)
- Search for related terms at **existing non-ontological resources**. These resources contain lexicons, thesauri, taxonomies and linked datasets. A critical part of this step is the identification of these resources. A good practice is to search at registries or lists of domain specific resources (*bottom-up*).
- Analysis of existing **raw data** specific to the domain (*e.g.* experimental data) that will have to be annotated using the model. The analysis of this data may result in new concepts that would be practically needed by the domain experts since they reflect real world requirements (*bottom-up*).

At the conceptualization phase it is advised to use Ontology Design Patterns (ODPs) [19] which facilitate the modeling of recurrent scenarios and provide guidelines for incorporating this knowledge into ontologies correctly.

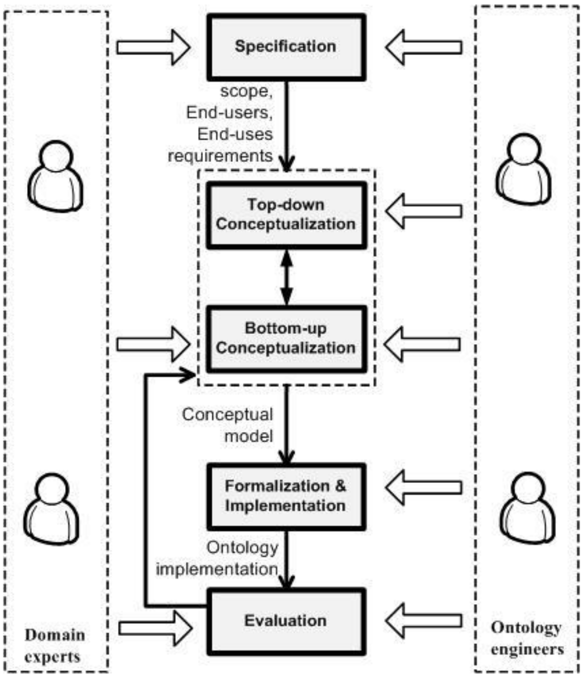


Fig. 2 A Collaborative methodology for building ontologies. Domain experts and ontology engineers collaborate in the specification, conceptualization and evaluation phases. The methodology phases flow from top to down.

The output of the conceptualization phase is a data model comprising of all identified concepts and relationships in a human-readable form (*e.g.* class diagram).

At the **implementation** phase the conceptual model is transformed into a computable model using an ontology language. An important decision to be taken is the implementation language to be used (*e.g.* OWL, RDF, RDFS). The language should cover the peculiarities of the model but must not be too complex by adding redundant complexity. Two activities have to be accomplished during the implementation:

- The **alignment** with other models and ontologies. This allows the definition of relationships between semantic models by identifying the matching concepts (*i.e.* have the same or similar meaning), thus enabling their interoperability. The semantic models to be used for the alignment are those detected at the conceptualization phase. Another key point is the language and property used to define the alignments (*e.g.* owl:sameAs, skos:closeMatch).
- The **re-use of upper ontologies,** where the model's concepts are mapped to concepts of an upper (top-level) ontology thus aiding the semantic integration across ontologies which are accessible "under" this upper ontology. The key here is the selection of the most appropriate upper ontology and the correct mapping of the model's concepts to the upper ontology concepts.

At the **evaluation** phase, we check if the developed semantic model fulfils the requirements defined in the specification phase (*e.g.* Does the model cover the end-uses? Does the model answer the competency questions?). Moreover we ensure that the model satisfies specific criteria, like the ones proposed for the evaluation of semantic models by existing methodologies [20-22]:

- *Lexicon & vocabulary*. Emphasizes the handling of concepts and the vocabulary used.
- *Hierarchy, Taxonomy*. Emphasizes taxonomic relations (is-a relations).
- *Semantic relations*. Evaluates other relations, which are not taxonomic relations.
- *Context or application*. Evaluates ontologies in their context of use/application.
- *Syntax*. Evaluates model conformity to syntactical requirements of formal language.
- *Structure and architecture*. Evaluates model conformity to predefined structural requirements.

Various methodologies for the evaluation of ontologies have been considered in the literature [23-26] depending on what kinds of ontologies are evaluated and for what purpose. The evaluation methodologies considered are the following:

- *Golden standard* [23]. Syntactic comparison between an ontology and a standard, which may be another ontology.
- *Application-based* [24]. Use of an ontology in an application followed by evaluation of the results.
- *Data or corpus driven* [25]. Comparison with a data source covered by the ontology.
- *Human assessment* [26]. Evaluation conducted by people based on criteria, like the ones presented in the previous paragraph, and patterns (*e.g.* use a class equivalence pattern to evaluate the equivalence relations defined at the ontology.)

The selection of the most appropriate evaluation methodology depends on the ontology/model that is evaluated, its intended uses and end-users. In the Evaluation phase the active involvement of the domain experts is required in order to guarantee that the resulting model does cover their needs. The output of the evaluation is used as feedback to the conceptualization phase in order to improve the model.

4. Case study: Developing the Cancer Chemoprevention semantic model

In this section we apply our methodology to create the Cancer Chemoprevention semantic model (CanCo). CanCo is one of the cornerstones of the GRANATUM project [27] that aims at bridging the information gap among biomedical researchers by offering homogenized access to resources needed to perform cancer chemoprevention experiments. In this context, CanCo is used in order to link the existing Life Sciences LOD Cloud by associating the concepts detected at the LOD Cloud with the concepts of CanCo [28]. This way the users are able to search across different data sources in a homogenized way by expressing their queries in CanCo terms.

4.1. Model Specification

As already stated in the Introduction, there is a need for a semantic model for Cancer Chemoprevention, because of the genericity of existing models and ontologies, which do not fully cover the peculiarities of the cancer chemoprevention domain.

The **end-users** of the model, which will be actually benefited from using it, are biomedical researchers, biologists and bioinformaticians. Representatives from these fields were actively involved in the model specification.

In order to define the **scope** of the model a questionnaire has been created and distributed aiming to collect the needs and expectations of the biomedical researchers. The questionnaire (available online at http://bit.ly/fZLh5K) contains 18 questions related to the kind of data the biomedical researchers use, problems faced when searching for data in different sources or when collaborating with other biomedical researchers etc. An extensive discussion on the questionnaire results has been conducted during a requirements collection workshop, where seven biomedical researchers and two ontology engineers participated.

As a result of the previous exercise, four main areas (Fig. 3) have been identified describing different aspects of cancer chemoprevention:

- The *Cancer chemoprevention* area enables the semantic annotation and representation of cancer chemoprevention related data and resources that define the main components of the chemoprevention procedure.
- The *Experimental representation* area facilitates the semantic annotation and representation of experimental data, procedures and protocols followed in order to identify and examine chemopreventive agents.
- The *Virtual screening* area facilitates the representation of data related to the execution of cancer chemoprevention experiments through computer simulation.
- The *Literature representation* area enables the semantic annotation and processing of scientific papers in online libraries related to cancer chemoprevention.

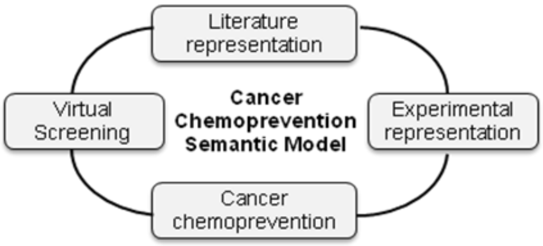

Literature representation
Virtual Screening
Cancer Chemoprevention Semantic Model
Experimental representation
Cancer chemoprevention

Fig. 3 Scope of the Cancer Chemoprevention semantic model

A set of four usage scenarios were co-designed with the biomedical experts [29]. The usage scenarios focus on the difficulties faced by biomedical researchers when evolving chemoprevention clinical trials design and planning, accelerate the conduction of the trials and improve the quality of the expected outcomes.

Based on the usage scenarios the **end-uses** of the model were defined. The users stated that the large amount of heterogeneous data sources related to cancer chemoprevention makes the search among them a cumbersome task. Moreover, there is a need to annotate experimental data and publications in order to be able to share and search for them. CanCo is used in order to facilitate the annotation of experimental data and publications related to cancer chemoprevention and unify the heterogeneous data sources allowing the query across them, using a common vocabulary (Fig. 1).

Throughout the specification phase (*i.e.* definition of end-users, end-uses, scope) the **model requirements** are formulated as competency questions. A total of 26 competency questions were created, most of them related to the function and characteristics of a chemopreventive agent, such as:

- What is the weight of a chemopreventive agent?
- In which pathways does a chemopreventive agent participate?
- Which are the natural and synthetic sources of a chemopreventive agent?
- In which assays is a chemopreventive agent examined?
- In which publications is a chemopreventive agent referenced?
- What does an experiment measure?
- Which is the process of an in-silico experiment?

4.2. Model Conceptualization

The first step of the conceptualization, as defined in the methodology, is the identification of the **core concepts** that came out of the specification and specifically the analysis of the competency questions. We manually examined the competency questions and the main concept identified is the Chemopreventive agent. This concept acts as a link between the four model areas. Other concepts identified are the pathway (where a chemopreventive agent may participate), the source (natural or synthetic) where a chemopreventive agent can be found in, the assay that may examine a chemopreventive agent and the publications that refer to an agent. These concepts act as a “seed” for the model to grow. In the next sections the top-down and bottom-up conceptualization of the model is clarified.

4.2.1. Top-down conceptualization

During the top-down conceptualization existing models (e.g. CancerGrid metamodel) and ontologies (e.g. ACGT) relevant to cancer chemoprevention were analyzed and clustered in order to identify the concepts and relationships relevant to CanCo.

In order to find these ontologies and models an extensive search, collaboratively with the domain experts, was conducted at the web and at biomedical related repositories, indicated by the biomedical researchers, such as BioPortal[1] and OBO Foundry[2]. Our search resulted in 18 ontologies. 5 of them (*i.e.* BiRO [30], CiTO [31], FaBiO [32], SIOC [33] and SWAN [34]) represent concepts related to the scientific literature and discourse (see Fig. 4 Scientific discourse ontologies), such as bibliographic records, citations, references and authors. 13 of them (*i.e.* ACGT [35], BioPAX [9], Biotop [36], CancerGrid Metamodel [37], EFO [6], GO [8], MeSH [38], MGED [39], NCI [40], OBI [7], RxNorm [41], UMLS [42], ISA [43]) are from the biomedical domain and represent concepts related to cancer chemoprevention, experimental representation and virtual screening (see Fig. 4 biomedical ontologies).

The analysis of existing models and ontologies comprised a multiphase iterative abstraction exercise, where their concepts were reviewed and compared with the core concepts identified at the specification phase. For example, one of the "seed" concepts is the Assay. This concept exists in the ISA framework which is based on three main concepts namely Investigation, Study, Assay. We then used the Investigation and Study concepts to extend CanCo.

The concepts of the ontologies and models were manually grouped in clusters with high similarity (only concepts related to cancer chemoprevention were encountered). In order to detect the similarity between concepts we checked their definition, name and synonyms. This means that the elements of a specific cluster were conceptually/semantically related despite differences in terminology (i.e. names). Then one representative concept from each cluster was extracted manually. For example, a cluster contains the concepts "clinical trial protocol", "protocol", "experiment design protocol", "study design", "trial protocol", "experimental design" and as a representative concept is selected the Protocol.

The biomedical researchers were actively involved in the top-down conceptualization by detecting concepts relevant to cancer chemoprevention and by supporting the grouping of the detected concepts in clusters with high similarity.

4.2.2. Bottom-up conceptualization

The bottom-up construction of the model identifies concepts based on existing linked datasets and experimental results that are relevant to cancer chemoprevention. More specifically, during the bottom-up conceptualization the following steps are followed:

- Analysis of publicly available datasets in the Linked Open Data Cloud tagged with "lifesciences" and/or "healthcare".
- Analysis of results obtained from cancer chemoprevention experiments that can be annotated with the CanCo semantic model.

The analysis of the **publicly available linked datasets** was based either on the data provided through the SPARQL endpoints of each dataset or through the searching mechanism provided by their Web site.

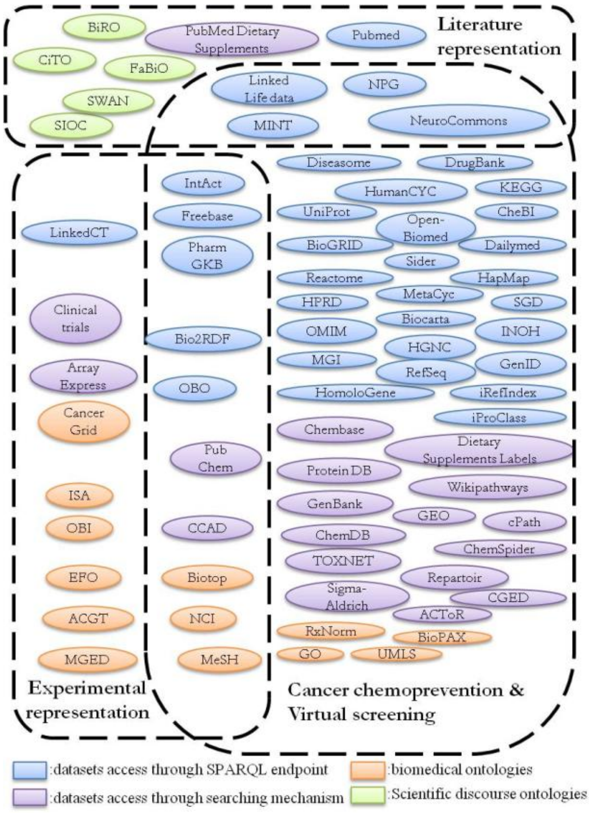


Fig. 4 A categorization of the surveyed ontologies and datasets grouped by the areas of CanCO.

[1] http://bioportal.bioontology.org/

[2] http://www.obofoundry.org/

In order to identify the linked datasets, a thorough search was conducted on the Web and in repositories containing biomedical-related SPARQL endpoints, such as BIO2RDF [3] and LinkedLifeData[4]. A total of 55 datasets were detected. 36 of them, i.e. ChEBI [44], Pubmed [45], DrugBank [46], KEGG [47], Reactome [48], UniProt [49], Diseasome [50], Dailymed [51], Sider [52], open-BioMed [53], BioGRID [54], Freebase [54], HapMap [55], HPRD [56], HumanCYC [57], IntAct [58], LinkedCT [59], MetaCyc [60], MINT [61], NeuroCommons [62], PharmGKB [63], NPG [64], OBO [65], Bio2RDF [66], LinkedLifeData [67], iProClass [68], HomoloGene [69], HGNC [70], Biocarta [71], INOH [72], GenID [73], OMIM [74], SGD [75], RefSeq [76], MGI [77] and iRefIndex [78], were accessed through a SPARQL endpoint - a single SPARQL endpoint may provide access to more than one dataset - (see Fig. 4 datasets accessed through SPARQL endpoint). 19 of the datasets, i.e. PubMed Dietary Supplement Subset [79], Dietary Supplements Labels Database [80], ClinicalTrials [81], TOXNET [82], ACToR [83], PubChem [84], Repartoire [85], CGED [86], ArrayExpress [87], GEO [88], GenBank [89], ChemSpider [90], Chembase [91], Sigma-Aldrich [92], ChemDB [93], CCAD [94], Wikipathways [95], cPath [96] and Protein DB [97], were accessed through the search mechanism available on their Web site (see Fig. 4 datasets accessed through searching mechanism).

The analysis of the linked datasets follows a two-step approach. First the elements of each dataset were reviewed, compared with the core concepts and clustered manually into semantically equivalent clusters. For each cluster, a representative concept was extracted. Moreover, representative attributes were reviewed, e.g. for the concept Molecule representative attributes are the “Formula”, “Molecular weight” and “Size” (see Fig. 5). The biomedical researchers were actively involved in the datasets analysis by detecting concepts relevant to cancer chemoprevention and by supporting the grouping of the detected concepts to clusters with high similarity. Second, an automatic analysis of the datasets is conducted. This analysis is based on the outcome (*i.e.* concept identified) of the first step and detects similar concepts taking into account their names, synonyms etc. A detailed description of the automatic analysis approach can be found in [28].

The **experimental data analysis** identified concepts by examining experimental data relevant to the cancer chemoprevention. For the experimental analysis two approaches were followed:

– Analyze experimental data provided by the biomedical researchers. For this reason we use two datasets, [98] and [99], that examine potential chemoprevention agents.
– Usually, experimental data is published in scientific publications. So, we searched through the LinkedLifeData dataset the number of Pubmed publications that mention concepts of the model. Only an approximation of the number of experimental datasets can be extracted based on this method, but it can derive the general trend. In order to achieve this, a separate SPARQL query was created for each concept. For example, the following SPARQL query counts the number of Pubmed publications that mention the concept “Chemopreventive agent” (umls-concept:C1516463).

```
Select (COUNT(?pub) as ?c)
 where {?pub rdf:type pubmed:Citation.
        ?pub lifeskim:mentions umls-concept:C1516463}
```

The results of the top-down and bottom-up conceptualization are presented in Table 1, which contains all the identified concepts. For each concept the table lists the ontologies/models (top-down) and linked datasets (bottom-up) that contain the specific concept (the name they use for that concept is presented in parenthesis). Moreover, the table reports if a concept is detected as part of the collaborative process (see User. Req.) or at the Experimental Data analysis (see Exp. Data).

4.2.3. Conceptual model

The outcome of the conceptualization phase is the CanCo model (Fig. 5) that comprises 27 concepts. The following paragraphs discuss these concepts in detail.

The core concept of the Cancer chemoprevention area is the *Chemopreventive agent*. A Chemopreventive agent is a *Natural* or *Synthetic* substance, such as a Drug or plant product that has shown some evidence of reducing the risk of development or recurrence of a tumor formation (i.e. *Cancer*) [40]. A Chemopreventive agent can prevent *Cancer* by interfering with a biological *Target* (e.g. nucleic acid, lipid, protein, sugar etc.) through a *Biological Mechanism* (e.g. anti-metastatic, anti-

[3] http://bio2rdf.org

[4] http://linkedlifedata.com/

proliferative, etc.). In other words, the Biological Mechanism is the way the Chemopreventive agent affects the *Target* in order to “break” the series of interactions that leads to a *Disease* (i.e. cancer). This series of interactions is captured by the *Pathway* which often forms a network that biologists have found useful to group together for organizational, historic, biophysical, or other reasons. Finally, the measurement of the *Toxicity* of a Chemopreventive agent is important, since it may cause injury to an organism in a dose-dependent manner.

The Experimental representation area is designed based on the ISA (Investigation – Study – Assay) framework [43] to capture data related to the experimental procedure. The main concept of the Experimental area is the *Study* that is a collection of *Assays* sharing the same *Protocol*. During a *Study*, *Measurements* are made based on a *Protocol*, which defines the followed procedure. The *Protocol* uses a set of *Experimental factors* that are the variable aspects of an experiment design (*e.g.* cell lines, organisms, biomaterial etc.) and can be documented separately in a *Published work*. A *Study* has an *Author* and is part of an *Investigation* that is a high-level concept to link related studies with the same subject. An *Assay* takes as input *Molecules* and investigates if they have chemopreventive action. Finally, *Assays* can be separated into *in-vivo* (performed on living organisms), *in-vitro* (performed outside of living organisms) and *in-silico* (performed on computer) based on the approach used.

The Virtual screening area defines concepts related to the execution of biomedical experiments through computer simulation. A type of *in-silico* Assay is the *Virtual Screening* that refers to computational technique used in drug discovery research. Each *in-silico* Assay uses a *Scientific Workflow* that is a pipeline of connected components (*in-silico* tools, models) exploited to perform an *in-silico* experiment.

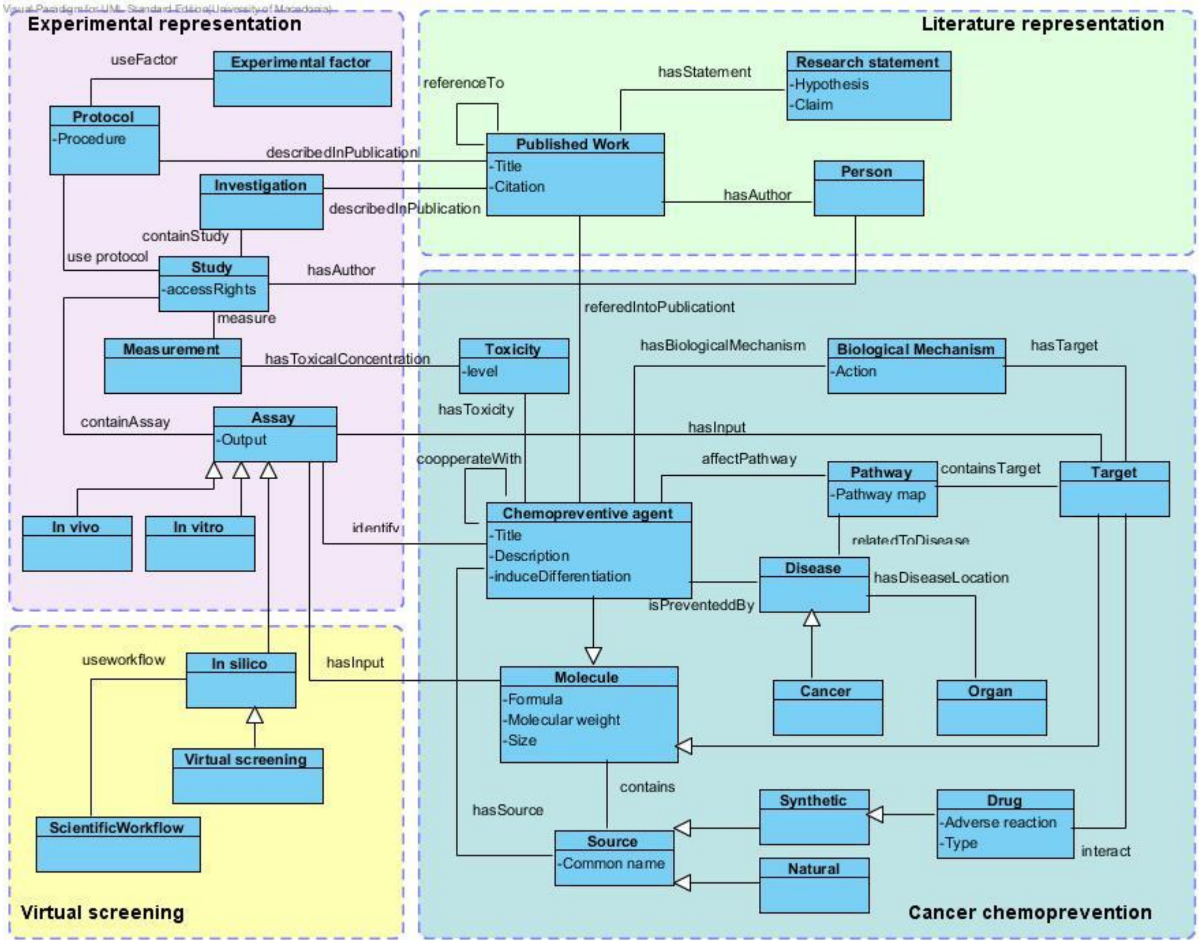


Fig. 5 The Cancer Chemoprevention model

	Concept	Top-down	Bottom-up		
		Ontology/Model	Linked Datasets	User Req.	Exp. data
Literature representation	Published Work	SWAN (Book, Journal, Newspaper article, Newspaper news, Web article), CiTO, BiRO, FaBiO (work)	PubMed,PubMed diet. sup, Neurocommon, NPG	☑	-
	Research statement	SWAN (Research statement), FaBiO (Expression)	-	☑	-
	Person	SWAN (agent), SIOC (user account)	PubMed, PubMed diet. Sup., NPG	-	-
Experimental representation	Experi-mental factor	ACGT(organism, substance sample), BIOTOP (organism part), EFO(experimental factor), MGED(experimental factor), OBI(organism) , NCI (organism, tissue), UMLS	PubChem, ArrayExpress	☑	☑
	Protocol	ACGT (clinical trial protocol), EFO (protocol), MGED (experiment design protocol), OBI (protocol, study design), CancerGrid (trial protocol), NCI (clinical trial protocol, experimental design),UMLS	PubChem, ArrayExpress	☑	☑
	Measurement	MGED, NCI	-	-	☑
	Investigation	NCI, ISA	-	-	☑
	Study	NCI, ISA	-	-	☑
	Assay	NCI, MeSH, ISA, EFO	Clinical trials, LinkedCT ArrayExpress,PubChem,	☑	☑
Virtual screening	Virtual screening	-	-	☑	-
	Scientific Workflow	NCI, MeSH	-	☑	☑
Cancer chemoprevention	Chemopreventive agent	NCI	LinkedLifeData, CCAD	☑	☑
	Toxicity	NCI, ACGT	TOXNET, ACToR	☑	☑
	Biological Mechanism	-	-	☑	-
	Pathway	NCI, BIOPAX	IntAct, PharmKGB, Wikipathways, KEGG, Repartoire, cPath, Reactome, MetaCYC, HapMap, Protein DB	☑	☑
	Target	ACGT(biological macromolecule), BIOPAX(protein, RNA, DNA), EFO (protein, DNA, RNA), OBI(macromolecule, nucleic acid, protein), GO (nucleic acid, protein) , NCI (nucleic acid, protein)	PharmKGB, Protein DB, Repartoire, GeneBank, UniProt, GEO, CGED, SigmaAldrich, HapMap BioGRID, HumanCYC, Open-biomed, MINT	☑	☑
	Disease	ACGT, OBI, NCI, MeSH, EFO (cancer), MGED (cancer)	PharmKGB, Diseasome, Repartoire, CGED	☑	☑
	Organ	ACGT, NCI	-	-	☑
	Molecule	BIOTOP(biological compound) EFO(chemical compound), MGED(compound), NCI(molecule)	ChEBI, Chembase, Chemspider, ChemDB	☑	☑
	Source	BIOPAX(biosource), NCI(source, natural source)	Diet. Sup. Labels,	☑	☑
	Drug	ACGT (Drug, chemotherapy drug), EFO, NCI (pharmaceutical substance), MGED, RxNorm	IntAct, DailyMed, Sider PharmKGB, DrugBank,	☑	☑

Table 1 The resources examined for the Cancer Chemoprevention semantic model (CanCo) conceptualization

The core concept of the Literature representation area is the *Published Work*. It refers to any type of publication that makes content publicly available (*e.g.* book, conference/journal article etc.). Each Published Work has at least one author that is a *Person*, and supports a number of *Research Statements*. The definition of Research Statement is based on the SWAN ontology [34] and is defined as a declarative sentence that has a hypotheses and a claim and is supported by a Published Work. The Published Work is an important concept for CanCo, since it may contain formal information for other concepts of the model (*e.g.* Protocols Chemopreventive agents).

The main modeling contribution of CanCo is the identification of the Chemopreventive agent as the main concept of the model and its correlation with concepts already defined in existing biomedical ontologies and linked datasets. More specifically, the Literature representation area contains the published information related to a Chemopreventive agent, the Experimental representation and the Virtual screening areas contain concepts for the representation of the experimental procedure followed in order to identify and examine a Chemopreventive agent. Finally, the Cancer chemoprevention area defines concepts that represent the way the Chemopreventive agent acts to prevent Cancer, as well as information about the Sources where an agent can be found.

At the conceptualization phase we considered the use of some basic ontology design patterns defined at OntologyDesignPatterns.org. Some indicative ontology design patterns used are: *i)* pattern corresponding to Datatype property, *ii)* class equivalence pattern *iii)* pattern corresponding to Object property. These patterns improve the ontological modeling, thus resulting to a more expressive and modular ontology.

4.3. Model Implementation

Until now the specification of CanCo remained at the conceptual (modeling) level. A machine-processable implementation of the model is required in order to *(i)* facilitate the model's uptake and reuse by the community, and *(ii)* utilize the model in the context of specific implementation. For this reason an implementation of CanCo in OWL lite was developed. OWL lite was selected as it is a well accepted and widely used Semantic Web standard that allows expressing relationship between concepts without introducing redundant complexity.

During the implementation, the classes and properties of the model (Fig. 5) where transformed into OWL classes and their relationships were encoded as OWL object properties. Fig. 6 shows an OWL representation of the Chemopreventive Agent.

An important part of the implementation phase is the model **alignment** that allows the definition of relationships with concepts of other ontologies that have the similar meaning. The ontologies that were used for alignment are those detected at the conceptualization phase. The alignment was semi-automatic and included two steps: *i)* for each cluster of similar concepts a relation is added between each of the cluster's concepts and the representative concept selected, *ii)* the concepts of CanCo are associated to concepts detected at the LOD Cloud using a specifically dedicated tool for the domain [28]. For example, the efo:protocol, acgt:clinical_trial_protocol, and the obi:study_design are linked to the CanCo:protocol. The selected property for the alignment is the skos:closeMatch because it defines "light" equivalence semantics compared with the strong equivalence semantics imposed by owl:sameAs.

In this context, CanCo is used in order to link the existing Life Sciences LOD Cloud by associating the concepts detected at the LOD Cloud with the concepts of CanCo [28]. This way the users are able to search across different data sources in a homogenized way by expressing their queries in CanCo terms

```
<owl:Class rdf:ID="ChemopreventiveAgent">
   <rdfs:subClassOf  rdf:resource="#Molecule"/>
   <rdfs:label> Chemopreventive Agent </rdfs:label>
  <rdfs:comment> A molecule  that can reduce the
       risk of developing tumor
  </rdfs:comment>
</owl:Class>
```

Fig. 6 OWL representation of the Chemopreventive Agent

CanCo is also linked with an upper ontology, Basic Formal Ontology (BFO) [100], that describes very general concepts that are the same across the biomedical domain. BFO was selected because it is a well structured ontology adopted by many biomedical ontologies, thus enabling the easy interoperability among them. In order to link CanCo with BFO, all the CanCo concepts are defined as subclasses of BFO concepts. The interested user can access the implementation of the CanCo ontology on BioPortal at http://bioportal.bioontology.org/ontologies/49087

4.4. Model Evaluation

We selected the Application-based methodology in order to evaluate the expressivity and completeness of CanCo in a real application, while the Human assessment methodology has been chosen in order to actively involve the biomedical researchers in the evaluation process. This way the adoption of the model by the biomedical community is facilitated. Moreover we adopted the use of OOPS! [101], which is a Web-based tool intended to detect potential errors, in order to improve the quality of CanCo. OOPS! detected a number of pitfalls (e.g. 21 missing annotations from ontology terms, 28 missing domain or range in properties etc.), which were used for refactoring CanCo. Some of the pitfalls could not be corrected, as they were related to imported ontologies (*i.e.* they are out of our control).

In order to simplify the human assessment evaluation a questionnaire was created (available online at http://bit.ly/HjXeeA). The questionnaire examined the completeness (*i.e.* Does the model answer the competency questions?), correctness (*i.e.* Does the model answer the competency questions correctly?), usability and the simplicity of CanCo. It was separated in two parts:

- The first part examines the usability and the simplicity of the model. In this part the biomedical researchers were asked to answer a tailored version of the System Usability Scale (SUS) [102] that is proposed by [103]. It contains 7 Likert scale questions (stating the degree of agreement or disagreement).
- The second part examines the correctness and the completeness of the model. It contains 4 questions related to the definitions of the model's concepts (in case no standard definitions are detected in existing ontologies) and 20 questions for the validation of the relations between the concepts that exist in CanCo. Moreover, it provides to the biomedical researchers the ability to express any disagreement or detect any concept or property missing.

The questionnaire was answered by seven biomedical experts (two lab directors, three researchers and two PhD students). Assuming the usability evaluation, the majority of the biomedical experts (71.42% agreement and 14.29% high agreement) declared that they could contribute to the model (Statement 1). This finding is related with the user's willingness to use and extend the model. The understanding of the model is examined by statements 2 and 6. A sufficiently large percentage of the biomedical experts (42.86%) found the model easy to understand (Statement 2). Moreover, most of the experts understand the conceptualization (Statement 6) of the model (71.42% agreement).

Regarding statements 3 and 5, the answers on the theoretical support needed by the users to understand the model vary; there are users that claim (Statement 3) that they do not need any support to understand the model (14.29%) while there are others that would need (28.57%). Most respondents believe that other biomedical experts would not understand the model easily. Finally, assuming the completeness (Statement 7) and integration (Statement 4) of the model, most of the users found the concepts of the model well integrated (71.42% agreement and 14.29% high agreement) and they believe that the model covers the needs of the cancer chemoprevention domain (42.86% agreement). The usability results are presented in detail in Table 2.

The questionnaire evaluates also the correctness and completeness of the model. The biomedical experts agreed with the concepts and properties of the model, but they also proposed changes to the definitions of the concepts as well as addition of new concepts and properties so that the model better described the cancer chemoprevention domain. For example the experts proposed the addition of new concepts such as the "Scientific workflow", "Toxicity" and "Biological mechanism". Moreover they proposed the addition of new properties such as the "affectPathway" which defines the pathways affected by a chemopreventive agent and the property "cooperateWith" which define that the chemopreventive agents may act in a co-operative mode. These changes were then provided as feedback to the conceptualization phase in order to make a top-down and bottom-up research based on the changes. The results of this procedure are depicted at Table 1 (*e.g.* at top-down conceptualization NCI, MeSH define the concept "Scientific workflow").

Assuming the Application-based evaluation methodology, an extension of Google Refine tool[5] has been developed. Google Refine is a tool for working with messy data and transforming it from one format into another. The extension created makes use of CanCo towards providing a user-friendly interface that biomedical researchers can use for extracting and annotating experimental data (e.g. in spreadsheet format) based on the CanCo semantic model. It is envisioned that users will likely further validate and improve the model through their interactions via the user interface.

[5] http://code.google.com/p/google-refine/

N	Evaluation statements	Strongly disagree	Disagree	Indifferent	Agree	Strongly agree
1	**I think that I could contribute to this model**	0.00%	0.00%	14.29%	71.42%	14.29%
2	**I find the model easy to understand**	0.00%	28.57%	28.57%	42.86%	0.00%
3	**I think that I would need further theoretical support to be able to understand this model**	14.29%	14.29%	28.57%	14.29%	28.57%
4	**I found the various concepts in this model were well integrated**	0.00%	0.00%	14.29%	71.42%	14.29%
5	**I would imagine that most biomedical experts would understand this model very quickly**	14.29%	28.57%	57.14%	0.00%	0.00%
6	**I am confident I understand the conceptualization of the model**	0.00%	0.00%	28.57%	71.42%	0.00%
7	**The concepts/properties of the model cover the needs of the Cancer Chemoprevention domain.**	0.00%	0.00%	57.14%	42.86%	0.00%

Table 2 Usability evaluation

5. Conclusion

In the ontology engineering literature there exist many methodologies for ontology creation that lack interaction with the end-users of the ontology, thus increasing the risk of creating an ontology that will not be useful or may not be accepted by the community of intended end-users. This paper proposes a collaborative methodology for developing ontologies where significant role in the methodology plays the feedback received from the domain experts at all development phases.

A limitation of the proposed methodology comes from the collaborative nature of the methodology, when the communication with the domain experts is difficult and/or the domain experts are unwilling to collaborate and return feedback.

Currently, there exist a large amount of data relevant to cancer chemoprevention, but they are spread across numerous heterogeneous data sources (ontologies, knowledge bases, linked datasets, databases etc.) Additionally, the existing vocabularies, ontologies and reference data in the literature are too generic and cannot cover the peculiarities of cancer chemoprevention. Therefore, we identified the need for a unified model for cancer chemoprevention that will enable the semantic annotation, sharing and interconnection of globally available cancer-chemoprevention-related and other types of biomedical resources.

In this work we utilized the proposed methodology to develop CanCo that provides a solution to the heterogeneity of the existing data sources and to the genericity of the available ontologies in the area of cancer chemoprevention. The model comprises four areas: *i)* Cancer chemoprevention *ii)* Experimental representation, *iii)* Virtual screening and *iv)* Literature representation. The main contributions of this work can be summarized as follows:

- It proposes a collaborative methodology for defining, developing and evaluating semantic models and ontologies. The novel part of the approach lies: *i)* in the adoption of a meet-in-the-middle approach where concepts emerged both in a bottom-up (*i.e.* analyzing the domain and interviewing the domain experts regarding their data needs) and top-down (*i.e.* analyze and integrate existing ontologies, vocabularies and data models) fashion *ii)* in the active engagement of the end-users during the actual development of the model and not just their limited involvement in the model evaluation.
- It defines the CanCo semantic model for the cancer chemoprevention domain. In this way it offers a common language in order to search and retrieve semantically-linked cancer chemoprevention related data and resources.

CanCo will be used in the GRANATUM FP7 project, in order to achieve interoperability and homogenized access of resources. In the context of the project, the model will drive the implementation of several tools, including the Google Refine extension mentioned earlier as well as a visual model editor that will allow biomedical researchers to easily extend the model by adding new concepts/properties in order to satisfy future individual requirements (*e.g.* annotation of more complex experimental data) not supported by the model. Finally, the end-users intend to use the model in order to facilitate their cancer chemoprevention studies by annotating and sharing experimental data and by searching for cancer chemoprevention related information across different data sources in a homogenized way.

Acknowledgements

This work is supported in part by the GRANATUM FP7 ICT Project under grant 270139.

References

[1] O. Corcho, M. Fernández-lópez, A. Gómez-pérez, and A. López, "Building legal ontologies with METHONTOLOGY and WebODE", V. R. Benjamins, P. Casanovas, J. Breuker, A. Gangemi (Eds.): *Law and the Semantic Web, LNCS volume 3369*, Springer-Verlag, Heidelberg, 2005, pp. 142-157.

[2] N. Noy and D. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology," *Tech. Rep. KSL-01-05 and SMI-2001-0800. Stanford, CA: Stanford University, Knowledge Systems Laboratory and Stanford Medical Informatics*, 2001.

[3] "Cancer prevetion network," *URL:* http://www.cancerpreventionnetwork.org/.

[4] A. Tsao, E. Kim, and W. K. Hong, "Chemoprevention of Cancer," *A Cancer Journal for Clinicians,* vol. 54, John Wiley & Sons, Ltd. 2004, pp. 150-180.

[5] R. C. Young and C. M. Wilson, "Cancer Prevention Past, Present, and Future," *Clinical Cancer Research,* vol. 8, pp. 11-16, 2002.

[6] J. Malone, E. Holloway, T. Adamusiak, M. Kapushesky, J. Zheng, N. Kolesnikov, A. Zhukova, A. Brazma, and H. Parkinson, "Modeling Sample Variables with an Experimental Factor Ontology," *Bioinformatics*, vol. 26, pp. 1112-1118, 2010.

[7] M. Courtot, W. Bug, F. Gibson, A. Lister, J. Malone, D. Schober, R. Brinkman, and A. Ruttenberg, "The OWL of Biomedical Investigations," C. Dolbear, A. Ruttenberg, U. Sattler (Eds.): OWLED *Workshop on OWL: Experiences and Directions, collocated with the 7th International Semantic Web Conference (ISWC-2008, CEUR-WS.org*, Karlsruhe, Germany, 2008.

[8] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, and J. T. Eppig, "Gene ontology: tool for the unification of biology," *The Gene Ontology Consortium. Nature Genet.*, vol. 25, pp. 25-29, 2000.

[9] G. D. Bader and M. P. Cary, "BioPAX – Biological Pathways Exchange Language " *Level 2, Version 1.0 Documentation*, doi:http://www.biopax.org/release/biopax-level2-documentation.pdf, 2005.

[10] M. Grüninger and M. Fox, "Methodology for the Design and Evaluation of Ontologies," in *IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing*, Montreal, Canada 1995.

[11] M. Uschold and M. Gruninger, "Ontologies: Principles, Methods and Applications," *Knowledge Engineering Review*, vol. 11, pp. 93-136, 1996.

[12] A. Öhgren and K. Sandkuhl, "Towards a methodology for ontology development in small and medium-sized enterprises", N. Guimarães, P. Isaias (Eds.): *International Conference on Applied Computing, IADIS*, Algarve, Portugal, 2005, pp 369-376.

[13] Z. Li, M. Yang, and K. Ramani, "A methodology for engineering ontology acquisition and validation," *Artif. Intell. Eng. Des. Anal. Manuf.*, vol. 23, pp. 37--51, 2009.

[14] M.C. Suárez-Figueroa. "NeOn Methodology for Building Ontology Networks: Specification, Scheduling and Reuse", *Dissertations in Artificial Intelligence*, vol. 338, ISBN: 978-3-89838-338-7. IOS Press, 2012.

[15] M.C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, A. Gangemi (eds.). "Ontology Engineering in a Networked World". ISBN: 978-3-642-24794-1. Springer, 2012.

[16] Y. Sure, S. Staab, and R. Studer, "On-To-Knowledge Methodology", S. Staab and R. Studer, (Eds.): *Handbook on Ontologies*, Springer, 2003, pp. 117-132.

[17] S. Pinto, S. Staab, and C. Tempich, "DILIGENT: Towards a fine-grained methodology for Distributed Loosely-controllled and evolvInG Engineering of oNTologies", R. L. de Mántaras, L. Saitta (Eds.): *Proceedings of the 16th Eureopean Conference on Artificial Intelligence* , IOS Press, 2004, pp 393-397.

[18] B. Villazón-Terrazas, L. M. Vilches-Blázquez, O. Corcho, and A. Gómez-Pérez, "Methodological Guidelines for Publishing Government Linked Data", D. Wood, (Ed.): *Linking Government Data*, New York: Springer 2011, pp. 27-49.

[19] V. Presutti, E. Blomqvist, E. Daga, A. Gangemi. "Pattern-Based Ontology Design", M.C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, A. Gangemi (Eds.): *Ontology Engineering in a*

Networked World (ISBN: 978-3-642-24794-1.), Springer Berlin, Heidelberg, 2012, pp. 35-64.

[20] M. B. Almeida, "A proposal to evaluate ontology content," *Applied Ontology*, vol. 4, pp. 245–265, 2009.

[21] G. Maiga and D. Williams, "A Flexible Approach for User Evaluation of Biomedical Ontologies," *International Journal of Computing and ICT Research*, vol. 2, pp. 62-74, 2008.

[22] M. Sabou and M. Fernandez, "Ontology (Network) Evaluation", M. C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, and A. Gangemi, (Eds.): *Ontology Engineering in a Networked World*, Springer Berlin, Heidelberg: 2012, pp. 193-212.

[23] A. Maedche and S. Staab, " Measuring similarity between ontologies", A. Gómez-Pérez, R. Benjamins (Eds.): *Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, 13th International Conference, EKAW 2002,* Lecture Notes in Computer Science 2473, Springer, 2002, pp 251-263.

[24] Y. Kalfoglou and B. Hu, "Issues with evaluating and using publicly available ontologies", C. Chen (Ed.): *Handbook of Software Engineering and Knowledge Engineering*, 2006.

[25] C. Brewster, H. Alani, S. Dasmahapatra, and Y. Wilk, " Data driven ontology evaluation," in *International Conference on Language Resources and Evaluation*, European Language Resources Association, Lisbon, Portugal, 2004.

[26] A. Gómez-Pérez, "Ontology evaluation", S. Staab and R. Studer, (Eds.): *Handbook on Ontologies*, Springer-Verlag, Berlin, 2004, pp. 251–274.

[27] "GRANATUM: A social collaborative working space semantically interlinking biomedical researchers, knowledge and data for the design and execution of in-silico models and experimments in cancer chemoprevention," *URL:* http://granatum.org/.

[28] A. Hasnain, R. Fox, S. Decker, and H. F. Deus, "Cataloguing and Linking Life Sciences LOD Cloud", *1st International Workshop on Ontology Engineering in a Data-driven World (OEDW 2012) collocated with 8th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2012)*, Galway, Ireland, 2012.

[29] GRANATUM project consortium, "Deliverable D1.1 - Requirements Analysis," *available at:* http://www.granatum.org/pub/bscw.cgi/d7070/Granatum_D1.1_Requirements_Analysis.pdf, 2011.

[30] "Bibliographic Reference Ontology (BiRO)," *URL:* http://purl.org/spar/biro.

[31] D. Shotton, "CiTO, the Citation Typing Ontology," *Journal of Biomedical Semantics*, vol. 1(Suppl 1):S6, 2010.

[32] "FRBR-aligned Bibliographic Ontology (FaBiO)," *URL:*http://purl.org/spar/fabio.

[33] U. Bojars, J. G. Breslin, V. Peristeras, G. Tummarello, and S. Decker, "Interlinking the Social Web with Semantics," *IEEE Intelligent Systems*, vol. 23, pp. 29-40, 2008.

[34] P. Ciccarese, E. Wu, G. Wong, M. Ocana, J. Kinoshita, A. Ruttenberg, and T. Clark, "The SWAN biomedical discourse ontology," *Journal of Biomedical Informatics*, vol. 41, pp. 739-751, 2008.

[35] M. Brochhausen, A. Spear, C. Cocos, G. Weiler, L. Martìn, A. Anguita, H. Stenzhorn, E. Daskalaki, F. Schera, U. Schwarz, S. Sfakianakis, S. Kiefer, M. Dörr, N. Graf, and M. Tsiknakis, "The ACGT Master Ontology and Its Applications - Towards an Ontology-Driven Cancer Research and Management System," *Journal of Biomedical Informatics*, vol. 44, pp. 8-25, 2011.

[36] E. Beißwanger, S. Schulz, H. Stenzhorn, and U. Hahn, "BioTop: An Upper Domain Ontology for the Life Sciences - A Description of its Current Structure, Contents, and Interfaces to OBO Ontologies," *Applied Ontology*, vol. 3, pp. 205-212, 2008.

[37] C. Crichton, J. Davies, J. Gibbons, S. Harris, A. Tsui, and J. Brenton, "Metadata-Driven Software for Clinical Trials," in *ICSE Workshop on Software Engineering in Health Care (SEHC '09)*, IEEE Computer Society Washington, DC, USA, 2009.

[38] H. J. Lowe and G. O. Barnett, "Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches," *Journal of the American Medical Assocation (JAMA)*, vol. 271, pp. 1103-1108, 1994.

[39] C. A. Ball and A. Brazma, "MGED standards: work in progress," *Omics 2006*, vol. 10, pp. 138-144, 2006.

[40] "National Cancer Institute (NCI) Thesaurus," http://ncit.nci.nih.gov/.

[41] S. Liu, M. Wei, R. Moore, V. Ganesan, and S. Nelson, "RxNorm: prescription for electronic

drug information exchange," *IT Professional*, vol. 7, pp. 17-23, 2005.

[42] D. Lindberg, B. Humphreys, and A. McCray, "The Unified Medical Language System," *Methods of Information and Medicine*, vol. 32, pp. 281-291, 1993.

[43] S.-A. Sansone, P. Rocca-Serra, D. Field, E. Maguire, C. Taylor, O. Hofmann, H. Fang, S. Neumann, W. Tong, L. Amaral-Zettler, K. Begley, T. Booth, L. Bougueleret, G. Burns, B. Chapman, T. Clark, L.-A. Coleman, J. Copeland, S. Das, A. d. Daruvar, P. d. Matos, I. Dix, S. Edmunds, C. T. Evelo, M. J. Forster, P. Gaudet, J. Gilbert, C. Goble, J. L. Griffin, D. Jacob, J. Kleinjans, L. Harland, K. Haug, H. Hermjakob, S. J. H. Sui, A. Laederach, S. Liang, S. Marshall, A. McGrath, E. Merrill, D. Reilly, M. Roux, C. E. Shamu, C. A. Shang, C. Steinbeck, A. Trefethen, B. Williams-Jones, K. Wolstencroft, I. Xenarios, and W. Hide, "Toward interoperable bioscience data," *Nature Genetics*, vol. 44, pp. 121–126, 2012.

[44] "Chemical Entities of Biological Interest (ChEBI)," *URL:http://www.ebi.ac.uk/chebi/*.

[45] "PubMed " *URL: http://www.ncbi.nlm.nih.gov/pubmed*.

[46] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Research*, vol. 36, pp. D901-D906, 2008.

[47] "Kyoto Encyclopedia of Genes and Genomes (KEGG)," *URL:http://www.genome.jp/kegg/*.

[48] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, and L. Matthews, et al. , "Reactome: a knowledgebase of biological pathways," *Nucleic Acids Research*, pp. D428-D432, 2005.

[49] "Universal Protein Resource (UniProt)," *URL:http://www.uniprot.org/*.

[50] "Diseasome," *URL:http://diseasome.eu/*.

[51] "Dailymed," *URL:http://dailymed.nlm.nih.gov*.

[52] "Sider," *URL:http://sideeffects.embl.de/*.

[53] "open-BioMed.org.uk," *URL:http://www.open-biomed.org.uk/*.

[54] "BioGRID," *URL:http://thebiogrid.org/*.

[55] "HapMap," *URL:http://hapmap.ncbi.nlm.nih.gov/*.

[56] K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. Somanathan, A. Sebastian, S. Rani, S. R. S, K. Harrys, S. Kanth, M. Ahmed, M. Kashyap, R. Mohmood, Y. Ramachandra, V. Krishna, B. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey, "Human Protein Reference Database " *Nucleic Acids Research*, vol. 37, pp. 767-72, 2009.

[57]P. Romero, J. Wagg, M. L. Green, D. Kaiser, M. Krummenacker, and P. D. Karp, "Computational prediction of human metabolic pathways from the complete human genome," *Genome Biology*, vol. 6, pp. 1-17, 2004.

[58] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, C. Jandrasits, R. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeiffenberger, P. Porras, A. Raghunath, B. Roechert, S. Orchard, and H. Hermjakob, "The IntAct molecular interaction database in 2012," *Nucleic Acids Research*, vol. 40 (Database issue), pp D841 - D846,2012.

[59] "Linked Clinical Trials (LinkedCT)," *URL:http://linkedct.org/*.

[60] R. Caspi, H. Foerster, C. A. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Y. Rhee, A. G. Shearer, C. Tissier, T. C. Walk, P. Zhang, and P. D. Karp, "The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases," *Nucleic Acids Res.*, vol. 36(Database issue), pp. 623–D631, 2008.

[61] A. Ceol, A. A. Chatr, L. Licata, D. Peluso, L. Briganti, L. Perfetto, L. Castagnoli, and G. Cesareni, "MINT, the molecular interaction database," *Nucleic Acids Res.*, vol. 38(Database issue), pp. 532 - 539, 2010.

[62] "NeuroCommons," *URL:http://neurocommons.org*.

[63] "Pharmacogenomics Knowledge Base (PharmGKB)," *URL:http://www.pharmgkb.org/*.

[64] "Nature Publishing Group: Linked Data Platform," *URL:http://data.nature.com*.

[65] B. Smith, M. Ashburner, C. Rosse, C. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S. A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis, "The OBO Foundry: coordinated evolu-

tion of ontologies to support biomedical data integration," *Nature Biotechnology*, vol. 25, pp. 1251 - 1255, 2007.

[66] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2RDF: towards a mashup to build bioinformatics knowledge systems," *J Biomed Inform*, vol. 41, pp. 706-716, 2008.

[67] Ontotext, "Linked Life Data," *URL:* http://linkedlifedata.com/.

[68] "Protein Information Resource:iProClass," *URL:*http://pir.georgetown.edu/iproclass/.

[69] NCBI, "HomoloGene," *URL:*http://www.ncbi.nlm.nih.gov/homologene.

[70] "HUGO Gene Nomenclature Committee (HGNC)," *URL:* http://www.genenames.org/.

[71] D. Nishimura, "BioCarta," *Biotech Software & Internet Report*, vol. 2, pp. 117-120, 2001.

[72] "INOX: Pathway Database," *URL:* http://www.inoh.org/.

[73] E. Blanco, G. Parra, and R. Guigó, "Using geneid to Identify Genes", D. Baxevanis and B. Davison, (Eds.): *Current Protocols in Bioinformatics*. vol. 18:4.3.1-4.3.28, John Wiley & Sons Inc., New York, 2007.

[74] NCBI, "Online Mendelian Inheritance in Man (OMIM)," *URL:* http://www.ncbi.nlm.nih.gov/omim.

[75] "Saccharomyces Genome Database (SGD)," *URL:* http://www.yeastgenome.org/.

[76] NCBI, "Reference Sequence (RefSeq)," *URL:* http://www.ncbi.nlm.nih.gov/RefSeq/.

[77] "Mouse Genome Informatics (MGI)," *URL:* http://www.informatics.jax.org/.

[78] "iRefIndex: A reference index for protein interaction data," *URL:* http://irefindex.uio.no.

[79] "PubMed Dietary Supplement Subset," *URL:*http://ods.od.nih.gov/research/PubMed_Dietary_Supplement_Subset.aspx.

[80] "Dietary Supplements Labels Database," *URL:*http://dietarysupplements.nlm.nih.gov/dietary/.

[81] "ClinicalTrials," *URL:*http://clinicaltrials.gov/.

[82] "TOXicology Data NETwork (TOXNET)," *URL:*http://toxnet.nlm.nih.gov/.

[83] "Aggregated Computational Toxicology Resource (ACToR)," *URL:*http://actor.epa.gov/actor/faces/ACToRHome.jsp.

[84] "PubChem," *URL:*http://pubchem.ncbi.nlm.nih.gov/.

[85] "Repartoire Database," *URL:*http://repairtoire.genesilico.pl/.

[86] "Cancer Gene Expression Database (CGED)," *URL:*http://lifesciencedb.jp/cged/.

[87] "ArrayExpress," *URL:*http://www.ebi.ac.uk/arrayexpress/.

[88] "Gene Expression Omnibus (GEO)," *URL:*http://www.ncbi.nlm.nih.gov/geo/.

[89] "GenBank," *URL:*http://www.ncbi.nlm.nih.gov/genbank/.

[90] "ChemSpider," *URL:*http://www.chemspider.com/.

[91] "Chemical Compounds Database (Chembase)," *URl:*http://www.chembase.com/.

[92] "Sigma-Aldrich," *URL:https://*www.sigmaaldrich.com/catalog/.

[93] "ChemDB," *URL:*http://cdb.ics.uci.edu/.

[94] D. Corpet and S. Tache, "Most effective colon cancer chemopreventive agents in rats: a systematic review of aberrant crypt foci and tumor data, ranked by potency," *Nutrition and Cancer*, vol. 43, pp. 1-21, 2002.

[95]A. Pico, T. Kelder, M. v. Iersel, K. Hanspers, B. Conklin, and C. Evelo, "WikiPathways: Pathway Editing for the People," *PLoS Biol*, vol. 6, e184 2008.

[96] E. G. Cerami, G. D. Bader, B. Gross, and C. Sander, "cPath: open source software for collecting, storing, and querying biological pathways," *BMC Bioinformatics*, vol. 7, p. 497, 2006.

[97] "Protein Database," *URL:*http://www.hprd.org/.

[98] R. G. Mehta, R. Naithani, L. Huma, M. Hawthorne, R. M. Moriarty, D. L. McCormick, V. E. Steele, and L. Kopelovich, "Efficacy of Chemopreventive Agents in Mouse Mammary Gland Organ Culture (MMOC) Model: A Comprehensive Review," *Current Medicinal Chemistry*, vol. 15, pp. 2785-2825, 2008.

[99] C. Gerhäuser, K. Klimo, E. Heiss, I. Neumann, A. Gamal-Eldeen, J. Knauft, G.-Y. Liu, S. Sitthimonchai, and N. Frank, " Mechanism-based in vitro screening of potential cancer chemopreventive agents," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 523-524, pp. 163-172, 2003.

[100] P. Grenon, "BFO in a Nutshell: A Bi-categorial Axiomatization of BFO and Comparison with DOLCE", *IFOMIS report*, University of Leipzig, 2003 available at:

http://www.ifomis.org/Research/IFOMISReports/IFOMIS%20Report%2006_2003.pdf

[101] M. Poveda-Villalón, M. C. Suárez-Figueroa, and A. Gómez-Pérez, "Validating Ontologies with OOPS!", A. Teije, J. Völker, S. Handschuh, H. Stuckenschmidt, M. d'Acquin, A. Nikolov, N. Aussenac-Gilles, and N. Hernandez, (Eds.): *Knowledge Engineering and Knowledge Management*. vol. 7603, Springer Berlin, Heidelberg, 2012, pp. 267-281.

[102] J. Brooke, "SUS: A "quick and dirty" usability scale", P. W. Jordan, B. Thomas, B. A. Weerdmeester, and I. L. McClelland (Eds.): *Usability evaluation in industry*, Taylor & Francis, London, 1996, pp. 189 -194.

[103] C. Nuria, "Ontology Evaluation through Usability Measures", R. Meersman, P. Herrero, and T. Dillon (Eds.): *Proceedings of OTM Workshops'2009*, LNCS 5872, 2009, pp. 594 - 603.