Collaborative development of a common semantic model for interlinking Cancer Chemoprevention linked data sources

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*
fistname.lastname@deri.org

Abstract. This paper proposes a hybrid collaborative methodology for creating a unified 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 interface for biomedical experts to search and retrieve linked cancer chemoprevention related data and Web resources. The model relies on widely known and adopted biomedical standards to represent: i) concepts from the literature, 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. The proposed methodology for the development of our model followed 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 and 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. Finally, we present a thorough evaluation of the model based on the feedback received from the domain experts.

Keywords: collaborative model development; Cancer Chemoprevention; Linked Data; HCLS

1. Introduction and motivation

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 [1]. It is considered as one of the most promising areas in current cancer research [2]. Data relevant to cancer chemoprevention is typically spread across a very large number of heterogeneous data sources, including ontologies, knowledge bases, databases with experimental results and publications.

As part of this work, we have analyzed approximately 70 biomedical data sources. We observed that most of them make use of different underlying schemas for knowledge representation. Although many of these employ semantically related conceptual elements with similar properties, they often use different identifiers and descriptions (e.g. "protocol" vs. "trial protocol") and data structures. These semantic incompatibilities hamper the uniform search across different sources as well as the integration of different data sources. They also increase the learning curve for the users, as users have to get accustomed to the specific vocabulary of every source, thus impeding the reuse of open bio-data.

The vocabularies, ontologies and reference data found in the literature are generic enough and do not fully cover the peculiarities of cancer chemoprevention. For example, the Experimental Factor Ontology (EFO) [3] and the Ontology for Biomedical Investi-

gations (OBI) [4] cover aspects related to the biomedical experiments but they do not connect the experiments to cancer chemoprevention processes. Moreover, the Gene Ontology (GO) [5] and BioPax [6] aim at standardizing the representation of genes and pathways, but they do not relate them with the chemoprevention action of an agent.

Therefore, a model that reuses and extends existing models, and interconnects biomedical data sources to facilitate the discovery of cancer-chemoprevention-related data is required.

In this vein, this work introduces a hybrid, collaborative methodology for creating the Cancer Chemoprevention Semantic Model (CanCO) that is designed to serve as a common model for the semantic annotation, sharing and interconnection of globally available cancer-chemoprevention-related resources. CanCO facilitates the delivery of machine-interpretable information regarding their structure and content, supporting the on demand discovery of published cancer chemoprevention related data.

CanCO provides a single interface for biomedical experts (i.e. biomedical researchers, biologists, clinicians, bioinformaticians and doctors) to search and retrieve linked cancer chemoprevention data and resources (Fig. 1).

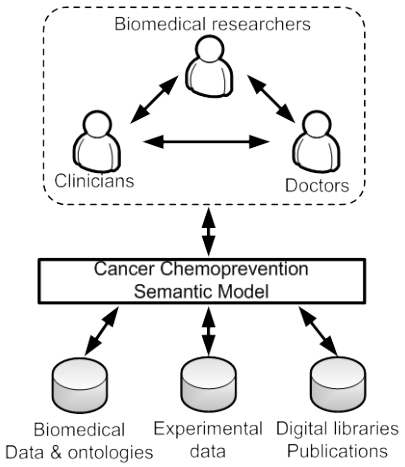

Fig. 1 The role of the Cancer Chemoprevention Semantic Model

The remainder of this paper is organized as follows. Section 2 presents the methodology that was followed for developing CanCO. Section 3 introduces CanCO and the methodology followed to identify the concepts of the model, as well as the encoding of the model in OWL. Section 4 discusses the evaluation of the model and a pilot application based on the model. Finally, in Section 5 we conclude the paper and discuss future research directions.

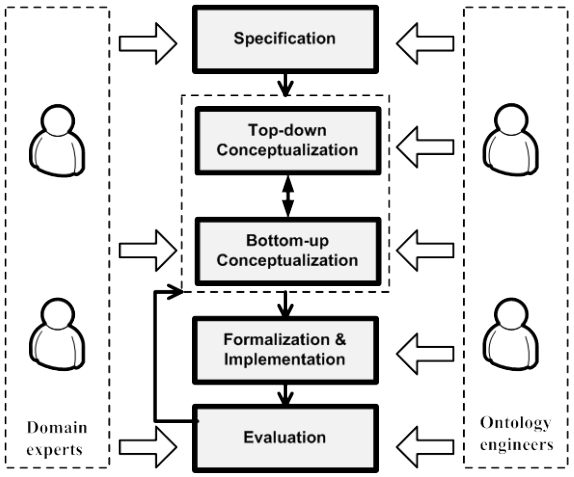

Fig. 2 Methodology for building CanCO

2. Methodology

The methodology followed for the definition of CanCO is based on a set of methodologies for ontology definition, namely METHONTOLOGY[7], Li et al.[8] and Öhgren et al. [9]. The novel part of the approach is the active engagement of the domain experts, i.e. the biologists, during the actual development of the model (specification and conceptualization) and not just their limited involvement in the model evaluation. The phases followed (Fig. 2.) are listed below:

- **Specification**. This phase investigates the reasons for which the semantic model is built and who the intended uses and the end-users are. At this stage the level of granularity of the concepts should also be taken into account. The specification of CanCO is described in Section 3.1 where the need for a unified cancer chemoprevention model is documented.
- **Conceptualization.** This phase identifies the concepts and relations of the model. The conceptualization of CanCO followed a "meet-in-the-middle" approach (Fig. 3). On the one hand relevant concepts emerged in a bottom-up fashion by analyzing the domain (i.e. existing data sets, user requirements and experimental data), on the other hand, it followed a top-down approach through analyses of existing ontologies and models. The result of the conceptualization activity is the conceptual model. The conceptualization of CanCO is described in Section 3.2.
- **Formalization and Implementation**. This phase transforms the conceptual model into a

formal or semi-computable model that later on can be translated into a computable model in any ontology language. The ontology language selected for the implementation is the Web Ontology Language (OWL). The Formalization and Implementation of CanCO is described in Section 3.3.

– **Evaluation.** This phase examines the completeness, correctness, usability and simplicity of CanCO through a human assessment evaluation based on a questionnaire. The feedback provided by the evaluation indicated corrections to the model, which were used to improve the Conceptualization step. The Evaluation of CanCO is described in Section 4.

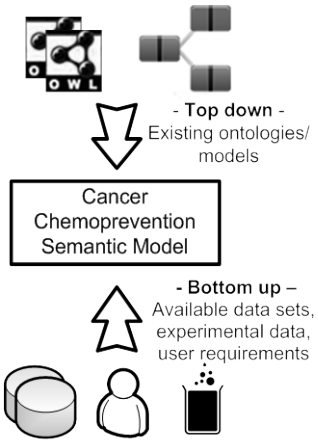

Fig. 3 Bottom-up and Top-down conceptualization of CanCO

3. Model design and development

3.1. Model Specification

As already stated in the Introduction, there is a need for a unified Cancer Chemoprevention Semantic Model. The reasons of this need can be summarized to the following:

– The heterogeneity of existing data sources relevant to cancer chemoprevention and the need to query them using a common vocabulary.
– The genericity of the existing ontologies that do not fully cover the peculiarities of cancer chemoprevention.

The model reflects the requirements of the biomedical experts that were actively involved in the model development. To do this, a questionnaire has been created and distributed aiming to detect the modeling needs and expectations of the biomedical experts for CanCO. An extensive discussion of the questionnaire results has been conducted during a requirements collection workshop which took place in May 2011. Eight biomedical experts and two ontology engineers participated in the workshop.

The model is separated into 4 spaces each describing a different aspect of cancer chemoprevention. The four spaces are connected through the main concept of the model that is the Chemopreventive Agent. The model spaces are the following:

– The *Cancer chemoprevention* space enables the semantic annotation and representation of cancer chemoprevention related data and resources. Specifically, it defines concepts and relationships that represent the way the chemopreventive agent acts in order to prevent a disease.
– The *Experimental representation* space facilitates the semantic annotation and representation of experimental data, procedures and protocols followed in order to identify and examine a chemopreventive agent.
– The *Virtual screening* space facilitates the representation of data related to the performance of cancer chemoprevention experiments through computer simulation.
– The *Literature representation* space enables the semantic annotation and processing of publications and scientific papers (in online libraries and digital archives) related to a chemopreventive agent, e.g. arguing for or against the cancer chemopreventive nature of a molecule.

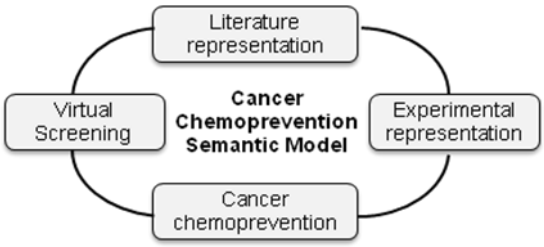

Fig. 4 Spaces of the Cancer Chemoprevention Semantic Model

3.2. Model Conceptualization

3.2.1. Top-down conceptualization

During the top-down conceptualization existing models and ontologies relevant to cancer chemoprevention were analyzed and clustered in order to retrieve the concepts and relationships relevant to CanCO (Fig. 4). More specifically, a total of 18 on-

tologies were detected after an extensive literature review. Five of them (BiRO [10], CiTO [11], FaBiO [12], SIOC [13] and SWAN [14]) represent concepts related to the scientific literature and discourse, such as bibliographic records, citations, references, authors etc, while 13 of them (ACGT [15], BioPAX [6], Biotop [16], CancerGrid Metamodel [17], EFO [3], GO [5], MeSH [18], MGED [19], NCI [20], OBI [4], RxNorm [21], UMLS [22], ISA [23]) are from the biomedical domain and represent concepts related to the Cancer Chemoprevention, the Experimental representation and the Virtual screening.

The analysis of existing models/ontologies comprised a multiphase iterative abstraction exercise, where the concepts of the models/ontologies were reviewed. The concepts of the models were manually grouped in clusters with high similarity (only concepts related to cancer chemoprevention were encountered). This means that the elements of a specific cluster were conceptually/semantically related despite differences in terminology. Then representative concepts from every cluster were extracted. For example a cluster contains the concepts "clinical trial protocol", "protocol", "experiment design protocol", "study design", "trial protocol", "experimental design" and as representative concept is selected the Protocol.

The results of the top-down conceptualization and the clustering are presented in Table 1. The concepts identified are grouped according to the model space they belong to. For each concept identified, the table lists the ontologies/models that contain the specific concept and in parenthesis the name they use for that concept.

3.2.2. Bottom-up conceptualization

The bottom-up construction of the model identifies concepts based on existing data sets and requirements that are relevant to the model spaces. 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" or "healthcare".
- Analysis of user requirements related to cancer chemoprevention obtained through interviews and feedback from the initial model.
- Analysis of results obtained from cancer chemoprevention experiments.

Publicly available datasets. The analysis of the publicly available 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. The analysis of the data sets is similar with the model analysis described in Section 3.2.1. The elements of each data set were reviewed and clustered manually into semantically equivalent clusters. For each cluster, a representative concept was extracted. Moreover, representative attributes were reviewed. For example for the concept Molecule representative attributes are the "Formula", "Molecular weight" and "Size" (see Fig. 6).

A total of 55 data sets were detected and analyzed after an extensive review of the state of the art. 36 of them, i.e. CheBI [24], Pubmed [25], DrugBank [26], KEGG [27], Reactome [28], UniProt [29], Diseasome [30], Dailymed [31], Sider [32], openBioMed [33], BioGRID [34], Freebase [34], HapMap [35], HPRD [36], HumanCYC [37], IntAct [38], LinkedCT [39], MetaCyc [40], MINT [41], NeuroCommons [42], PharmGKB [43], NPG [44], OBO [45], Bio2RDF [46], LinkedLifeData [47], iProClass [48], HomoloGene [49], HGNC [50], Biocarta [51], INOH [52], GenID [53], OMIM [54], SGD [55], RefSeq [56], MGI [57], iRefIndex [58] were accessed through a SPARQL endpoint (a single SPARQL endpoint may provide access to more than one data set), while 19 of them, i.e. PubMed Dietary Supplement Subset [59], Dietary Supplements Labels Database [60], ClinicalTrials [61], TOXNET [62], ACToR [63], PubChem [64], Repartoire [65], CGED [66], ArrayExpress [67], GEO [68], GenBank [69], ChemSpider [70], Chembase [71], Sigma-Aldrich [72], ChemDB [73], CCAD [74], Wikipathways [75], cPath [76], Protein DB [77], were accessed through the search mechanism available on their Web site.

Analysis of user requirements. The analysis of the user requirements was based on structured questionnaires and on relevant usage scenarios:

- A questionnaire[1] has been employed in order to elicit requirements from the biomedical experts. It contains 18 questions related to the kind of data biomedical experts use, problems faced when searching for data in different sources or when collaborating with other biomedical experts etc.
- A set of 4 usage scenarios were designed in a collaboration fashion with the biomedical experts [78]. The usage scenarios focus on the dif-

[1] http://bit.ly/fZLh5K

ficulties 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.

– Finally, a questionnaire [2] was used for the evaluation of the model. The biomedical experts were asked to evaluate the completeness and correctness of the produced model. The feedback provided by the evaluation may indicate corrections to the model, which will then be used to improve the Conceptualization step.

Experimental data analysis. The experimental data analysis identified concepts by examining experimental data relevant to the cancer chemoprevention. To succeed this, two sets of experimental data were analyzed. The first experimental data set [79] examines the activity of more than 200 synthetic and natural product-derived molecules and identifies potential chemoprevention agents. The second dataset [80] identifies potential cancer chemopreventive constituents. A number of known chemopreventive substances have been tested belonging to several structural classes as reference compounds for the identification of novel chemopreventive agents or mechanisms.

The results of the bottom-up conceptualization are presented in Table 1. The table contains the concepts identified, for each concept identified the table lists the data sets that contain the specific concept. Moreover the table reports if a concept is detected at the User Requirements (see User Req.) or the Experimental Data (see Exp. Data) analysis.

The analysis of the last two sections examined a number of resources (ontologies and datasets) related to cancer chemoprevention. Fig. 5 summarizes these resources clustered based on the spaces of the model. There exist overlaps between the spaces since many resources cover more than one space.

3.2.3. Conceptual model

The concepts identified by both approaches (top-down and bottom-up) were then merged in CanCO depicted in Fig. 6. The final model comprises of 27 concepts distributed across the four spaces of the model, namely Cancer chemoprevention, Experimental representation, Virtual screening and Literature representation. The following paragraphs discuss these concepts in detail.

[2] http://bit.ly/HjXeeA

The core concept of the Cancer chemoprevention space 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 that it may reduce the risk of developing or recurrence of tumor formation (i.e. *Cancer*) [20]. 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-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.

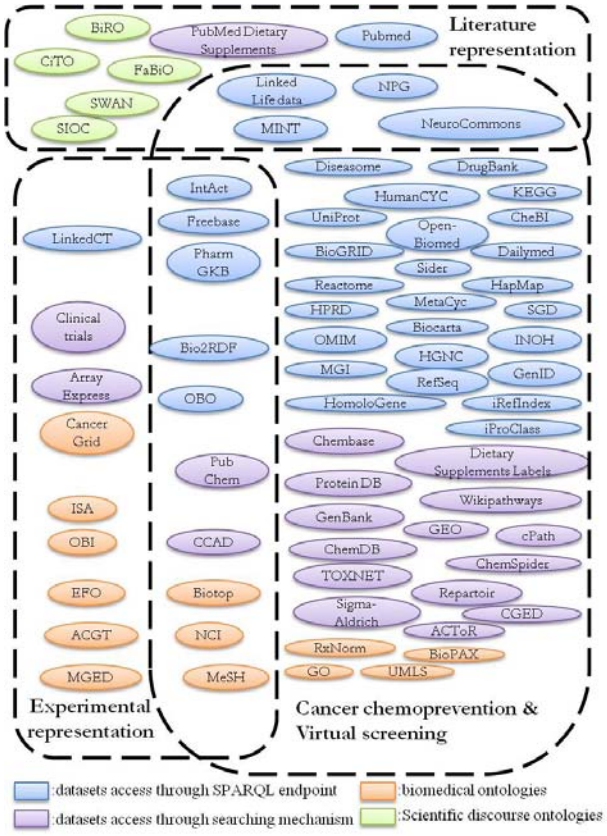

Fig. 5 Categorization of the Ontologies/datasets based on the spaces of the Cancer Chemoprevention Semantic Mode (CanCO)

	Concept	Top-down	Bottom-up		
		Ontology	Data sets	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	Experimental 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	-	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 Cancer Chemoprevention Semantic Model (CanCO) conceptualization

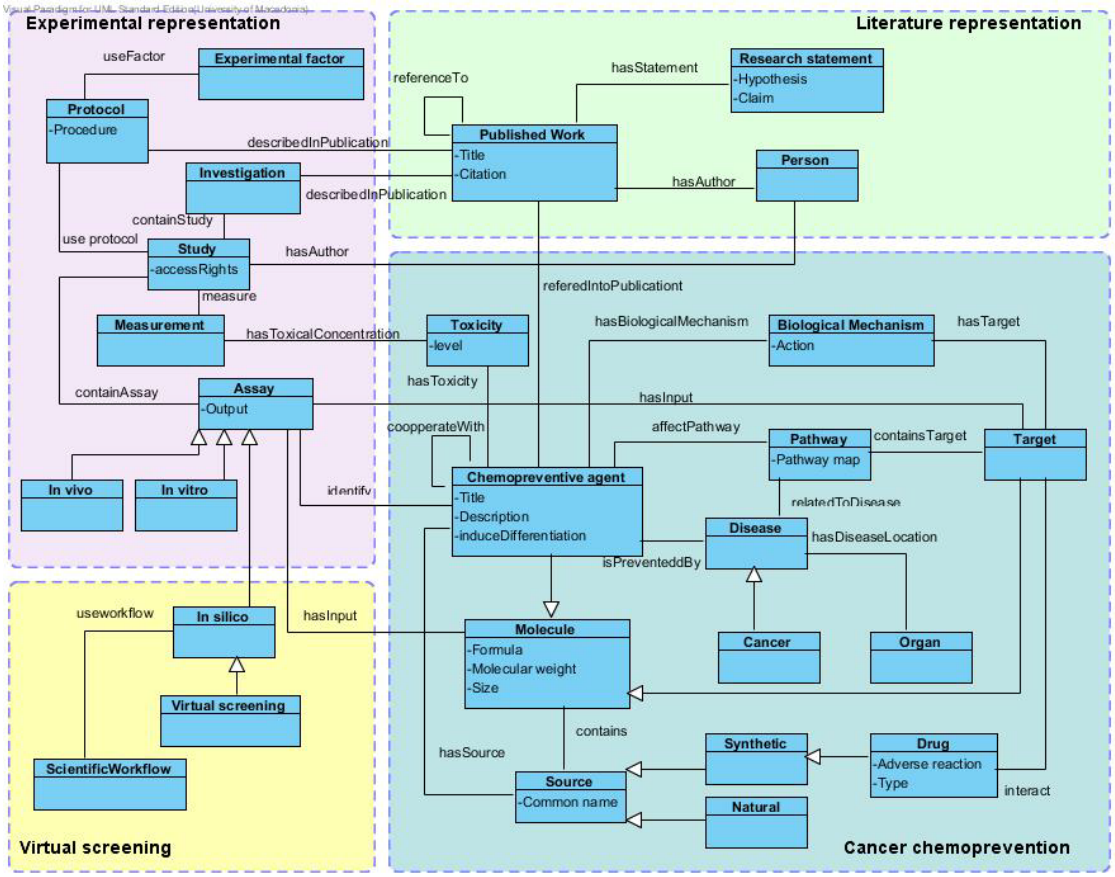

Fig. 6 The Cancer Chemoprevention Semantic Model

The Experimental representation space is designed based on the ISA (Investigation – Study – Assay) framework [23] to capture data related to the experimental procedure. The main concept of the Experimental space 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 procedure followed. 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* (done in living organisms), *in-vitro* (done outside of living organisms) and *in-silico* (performed on computer) based on the methodology used.

The Virtual screening space 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.

The core concept of the Literature representation space is the *Published Work*. It refers to any type of publication that makes content available to public (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 [14] 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 and documentation for other concepts of the model (e.g. Protocols, Investigations and 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 data sets. More specifically the Literature representation space contains the published information related to a Chemopreventive agent, the Experimental representation and Virtual screening spaces contain concepts for the representation of the experimental procedure followed in order to identify and examine a Chemopreventive agent. Finally, the Cancer chemoprevention space defines concepts that represent the way the Chemopreventive agent acts in order to prevent Cancer, as well as information about the Sources that a Chemopreventive agent can be found to.

3.3. Model Formalization and Implementation

The formalization of the model transforms the conceptual model into a formal or semi-computable model. For the formalization of CanCO a standard template is used to formally define the concepts and the properties of the model. The template contains information like the name, a universal resource identifier (URI), the definition and the is-a relations of the class/property. Table 2 contains a formalization of the Chemopreventive Agent.

ChemopreventiveAgent	
URI	../CanCO/ns#ChemopreventiveAgent
Definition	A Chemopreventive agent is a Natural or Synthetic substance, such as a Drug, or plant product, that has shown some evidence that it may reduce the risk of developing or recurrence of tumor formation.
is-a	Molecule

Table 2 Formalization of the Chemopreventive Agent

Until now the specification of CanCO remained at the conceptual (modeling) level. A machine-processable implementation of the model was required in order to *(i)* facilitate the model's uptake and reuse by the community, and *(ii)* use the model in the context of specific implementation. Therefore, an implementation of CanCO in OWL was developed. OWL was selected as it is a well accepted and widely used Semantic Web standard that allows expressing complex relationship between concepts.

During the implementation, the classes and properties of the model (Fig. 6) where transformed into OWL classes and their relationships were encoded as OWL object properties. Table 3 shows an OWL representation of the Chemopreventive Agent.

CanCO is linked with two biomedical ontologies, Experimental Factor Ontology (EFO) and Advancing Clinico-Genomic Trials on Cancer (ACGT), since they are highly-related with cancer chemoprevention. These ontologies were imported by CanCO and similar concept where linked, for example efo:Disease and acgt:Disease are linked with the Disease concept defined by CanCO. CanCO is also linked with an upper ontology, Basic Formal Ontology (BFO) [81] that describes very general concepts that are the same across the biomedical domain. For example the concept Assay is defined as a sub-concept of bfo:Process. Thus, it enables a semantic interoperability between a large number of ontologies which are accessible ranking "under" this upper ontology. The working draft of the ontology can be accessed at Bioportal[3].

```
<owl:Class rdf:ID="ChemopreventiveAgent">
  <rdfs:subClassOf rdf:resource="#Molecule"/>
  <rdfs:label> Chemopreventive Agent </rdfs:label>
  <rdfs:comment> A Chemopreventive agent is a
      Natural or Synthetic substance, such as a Drug,
      or plant product, that has shown some evidence
      that it may reduce the risk of developing or re-
      currence of tumor formation.
  </rdfs:comment>
</owl:Class>
```

Table 3 OWL representation of the Chemopreventive Agent

4. Demonstration and Evaluation of the model

CanCO needs to be evaluated and tested according to specified criteria. These criteria are proposed in existing methodologies for model evaluation [82, 83] and are:

- *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 model 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.

[3] http://bioportal.bioontology.org/ontologies/3030

N	Question	High disagree	Disagreement	Indifferent	Agreement	High 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 4 Usability evaluation

Various methodologies for the evaluation of ontologies have been considered in the literature, depending on the ontologies and the evaluation purpose. The evaluation methodologies adopted are:

- *Application Based* [84]. Use of an ontology in an application followed by evaluation of the results.
- *Human assessment*[85]. Evaluation conducted by people based on criteria and patterns.

The Application Based methodology is selected 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 experts in the evaluation process. In this way the adoption of the model by the biomedical community is facilitated.

In order to simplify the human assessment evaluation a questionnaire was created[4]. The questionnaire examined the completeness, correctness, usability and the simplicity of CanCO. It was separated into two parts:

- The first part examines the usability and the simplicity of the model. In this part the biomedical experts (i.e. biomedical researchers, biologists, clinicians and doctors) were asked to answer a tailored version of the System Usability Scale (SUS) [86] that is proposed by [87] in order to evaluate the understanding and agreement of the biomedical experts regarding CanCO as a whole. 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 the biomedical experts the ability to express any disagreement or detect any concept or property missing.

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 (Question 1). This finding is related with the user's willingness to use and extend the model. The understanding of the model is examined by Questions 2 and 6. Most of the biomedical experts (42.86%) found the model easy to understand (Question 2). Moreover, most of the experts understand the conceptualization (Question 6) of the model (71.42% agreement).

Regarding questions 3 and 5, the answers vary about the theoretical support needed by the users to understand the model. Finally, assuming the completeness (Question 7) and integration (Question 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 4.

Except from the usability evaluation, the questionnaire checks also the correctness and completeness of the model. The biomedical experts agreed with the

[4] http://bit.ly/HjXeeA

concepts and properties of the model but they also proposed changes to the definitions of the concepts as well as addition/deletion of concepts properties. These changes were incorporated into CanCO (Fig. 6). The concepts that derived from the evaluation process (e.g. Scientific workflow, Toxicity, Biological mechanism) are marked to be derived from the User requirements as part of the bottom-up conceptualization (see Table 1).

Assuming the Application-based evaluation, an extension of Google Refine tools[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 experimental researchers can easily understand and use for submitting data. It is envisioned that users will likely further validate and improve the model through their interactions via the user interface.

5. Conclusion

Currently, there exist a big number of data relevant to cancer chemoprevention but they are spread across a very large number of heterogeneous data sources (ontologies, knowledge bases, databases with experimental results and publications). Additionally, the existing vocabularies, ontologies and reference data in the literature are generic enough 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 proposed the Cancer Chemoprevention Semantic Model (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 of four spaces namely: *(i)* Cancer chemoprevention *(ii)* Experimental representation, *(iii)* Virtual screening and *(iv)* Literature representation. The methodology proposed for the development of the model follows a "meet-in-the-middle" approach, where the concepts emerge both in a bottom-up (analysis of the domain) and top-down (analysis of existing models/ontologies) fashion. Significant role in the methodology plays the feedback received from the domain experts at different phases of the development. The main contributions of this work can be summarized in the following:

- It proposes and realizes a hybrid collaborative methodology for defining, developing and evaluating the Cancer Chemoprevention Semantic Model. The novel part of the approach was on crowdsourcing the model and asking for feedback from the biomedical experts during each phase.
- It defines a unified model for the cancer chemoprevention domain. In this way it offers a common language for the biomedical experts in order to search and retrieve semantically-linked cancer chemoprevention related data and resources.
- It lowers the semantic interoperability barriers and thus contributes to the reusability of existing biomedical ontologies and data described using different semantic models.
- It identifies the Chemopreventive agent as the main concept of the model and correlates it with concepts already defined by other ontologies or data sets.

As part of our future research, we plan to exploit CanCO in the GRANATUM project [88] that aims at bridging the information gap among biomedical researchers by offering homogenized access to resources needed to perform cancer chemoprevention experiments and conduct studies on large-scale datasets. The model will be one of the pillars on which the GRANATUM approach will build 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 biologists to easily create extensions of the model in order to satisfy their individual requirements.

Acknowlegements

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

References

[1] A. Tsao, E. Kim, and W. K. Hong, "Chemoprevention of Cancer," *A Cancer Journal for Clinicians*, vol. 54, pp. 150-180, 2005.

[2] R. C. Young and C. M. Wilson, "Cancer Prevention Past, Present, and Future," *Clini-*

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

cal Cancer Research, vol. 8, pp. 11-16, 2002.

[3] 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.

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

[5] 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.

[6] 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.

[7] O. Corcho, M. Fernández-lópez, A. Gómez-pérez, and A. López, "Building legal ontologies with METHONTOLOGY and WebODE," in *Law and the Semantic Web, number 3369 in LNAI*: Springer-Verlag, 2005, pp. 142--157.

[8] 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.

[9] A. Öhgren and K. Sandkuhl, "Towards a methodology for ontology development in small and medium-sized enterprises," in *IADIS International Conference on Applied Computing*, 2005, pp. 369 - 376.

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

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

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] C. Crichton, J. Davies, J. Gibbons, S. Harris, A. Tsui, and J. Brenton, "Metadata-Driven Software for Clinical Trials," *Proceedings of the 2009 ICSE Workshop on Software Engineering in Health Care (SEHC '09)*, 2009.

[18] 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.

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

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

[21] 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.

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

[23] 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. Ma-

tos, 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.

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

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

[26] 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.

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

[28] 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.

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

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

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

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

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

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

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

[36] 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.

[37] 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.

[38] 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. [Epub ahead of print], 2011.

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

[40] 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.

[41] 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.

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

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

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

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

[46] 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.
[47] Ontotext, "Linked Life Data," *URL:* http://linkedlifedata.com/.
[48] "Protein Information Resource:iProClass," *URL:*http://pir.georgetown.edu/iproclass/.
[49] NCBI, "HomoloGene," *URL:*http://www.ncbi.nlm.nih.gov/homologene.
[50] "HUGO Gene Nomenclature Committee (HGNC)," *URL:* http://www.genenames.org/.
[51] D. Nishimura, "BioCarta," *Biotech Software & Internet Report*, vol. 2, pp. 117-120, 2001.
[52] "INOX: Pathway Database," *URL:* http://www.inoh.org/.
[53] E. Blanco, G. Parra, and R. Guigó, "Using geneid to Identify Genes," in *Current Protocols in Bioinformatics*. vol. 1, D. Baxevanis and D. B. Davison, Eds. New York: John Wiley & Sons Inc., 2002.
[54] NCBI, "Online Mendelian Inheritance in Man (OMIM)," *URL:* http://www.ncbi.nlm.nih.gov/omim.
[55] "Saccharomyces Genome Database (SGD)," *URL:* http://www.yeastgenome.org/.
[56] NCBI, "Reference Sequence (RefSeq)," *URL:* http://www.ncbi.nlm.nih.gov/RefSeq/.
[57] "Mouse Genome Informatics (MGI)," *URL:* http://www.informatics.jax.org/.
[58] "iRefIndex: A reference index for protein interaction data," *URL:* http://irefindex.uio.no.
[59] "PubMed Dietary Supplement Subset," *URL:*http://ods.od.nih.gov/research/PubMed_Dietary_Supplement_Subset.aspx.
[60] "Dietary Supplements Labels Database," *URL:*http://dietarysupplements.nlm.nih.gov/dietary/.
[61] "ClinicalTrials," *URL:*http://clinicaltrials.gov/.
[62] "TOXicology Data NETwork (TOXNET)," *URL:*http://toxnet.nlm.nih.gov/.
[63] "Aggregated Computational Toxicology Resource (ACToR)," *URL:*http://actor.epa.gov/actor/faces/ACToRHome.jsp.
[64] "PubChem," *URL:*http://pubchem.ncbi.nlm.nih.gov/.
[65] "Repartoire Database," *URL:*http://repairtoire.genesilico.pl/.
[66] "Cancer Gene Expression Database (CGED)," *URL:*http://lifesciencedb.jp/cged/.
[67] "ArrayExpress," *URL:*http://www.ebi.ac.uk/arrayexpress/.
[68] "Gene Expression Omnibus (GEO)," *URL:*http://www.ncbi.nlm.nih.gov/geo/.
[69] "GenBank," *URL:*http://www.ncbi.nlm.nih.gov/genbank/.
[70] "ChemSpider," *URL:*http://www.chemspider.com/.
[71] "Chemical Compounds Database (Chembase)," *URl:*http://www.chembase.com/.
[72] "Sigma-Aldrich," *URL:https://*www.sigmaaldrich.com/catalog/.
[73] "ChemDB," *URL:*http://cdb.ics.uci.edu/.
[74] 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.
[75] A. Pico, T. Kelder, M. v. Iersel, K. Hanspers, B. Conklin, and C. Evelo, "WikiPathways: Pathway Editing for the People," *PLoS Biol, doi:10.1371/journal.pbio.0060184*, vol. 6, 2008.
[76] 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.
[77] "Protein Database," *URL:*http://www.hprd.org/.
[78] "GRANATUM project: Deliverable D1.1 - Requirements Analysis," 2011.
[79] 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,.
[80] 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.
[81] P. Grenon, "BFO in a Nutshell: A Bicategorial Axiomatization of BFO and

Comparison with DOLCE," *Institute for Formal Ontology and Medical Information Science (IFOMIS)*, 2003.

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

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

[84] Y. Kalfoglou and B. Hu, "Issues with evaluating and using publicly available ontologies," in *In 4th International Workshop on Evaluation of Ontologies for the Web (EON 2006) at the 15th International World Wide Web Conference* Edinburgh, UK, 2006.

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

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

[87] C. Nuria, "Ontology Evaluation through Usability Measures," in *Proceedings of OTM Workshops'2009*, Vilamoura, Portugal, 2009, pp. 594 - 603.

[88] "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/*.