Semantics Annotations of Ontology for Scenario: Anthropogenic Impact and Climate Change Issues

Miroslav Kubásek, Jiří Hřebíček, Jiří Kalina, Ladislav Dušek, Ivan Holoubek

Semantics Annotations of Ontology for Scenario: Anthropogenic Impact and Climate Change Issues

Miroslav Kubasek[1], Jiri Hrebicek[1], Jiri Kalina[1], Ladislav Dusek[1], Ivan Holoubek[2],

[1]IBA, Masaryk University, Kamenice 126/3, 625 00 Brno, Czech Republic
{kubasek, hrebicek, urbanek, dusek}@iba.muni.cz
[2]RECETOX, Masaryk University, Kamenice 126/3, 625 00 Brno, Czech Republic
holoubek@recetox.muni.cz

Abstract. The synthesis of existing Persistent Organic Pollutants (POPs) pollution monitoring databases with epidemiological data is considered for identifying some impacts of POPs on human health. This task requires new, rich, data, services and models discovery capabilities from a multitude of monitoring networks and web resources. The FP7 project TaToo (Tagging Tool based on a Semantic Discovery Framework) is setting up a semantic web solution to close the discovery gap that prevents a full and easy access to web resources. The use of TaToo tools together with software GENASIS and SVOD is discussed as TaToo validation scenario for anthropogenic impact and global climate change influence on POPs trajectory. This paper contains the first propose of POPs and cancer domain ontology intended for TaToo framework.

Keywords: TaToo, semantic web, POPs, Stockholm Convention, GENASIS, SVOD, anthropogenic impact, global climate change.

1 Introduction

The FP7 project TaToo (Tagging Tool based on a Semantic Discovery Framework) aims to set up a semantic web solution to close the discovery gap that prevents a full and easy access to environmental resources on the web [9]. The core of the project will focus on the development of tools allowing third parties to easily discover environmental resources (data and/or services residing on different information nodes) on the web and to add valuable information in the form of semantic annotations to these resources, thus facilitating future usage and discovery, and kicking off a beneficial cycle of information enrichment.

TaToo validates the usability of its developments through the implementation of three different validation scenarios. All three scenarios are embedded in highly complex environmental domains and are therefore mainly addressed to domain expert groups and communities as well as to technically skilled users.

This paper contains the detailed framework of the Masaryk university[1] validation scenario (MU scenario), which is the anthropogenic impact and the influence of

[1] http://www.muni.cz

global climate change on this impact [5], [6]. The purpose of the paper is to give the overview of the MU scenario (background, objectives, and available tools), the definition of users who will use the TaToo tools, describe the possible use cases, and to provide with the mock ups.

2 Description of the scenario

The MU scenario of TaToo project named *Anthropogenic impact and global climate change* is managed by Masaryk University (MU). This scenario is dealing with the correlation of environmental pollutants and their health impact on the population and the correlation of transport of environmental pollutants with global climate change [5], [6]. The aim is to create a central place for researchers, domain experts and decision makers to discover and access interdisciplinary knowledge in more efficient and usable way that is the currently state of the art. Due to the fact that there is an enormous amount of information resources in scientific fields, which is steadily growing, available search mechanisms like search engines, scientific networks and similar technologies are not sufficient to meet the complex requirements of today's researchers and scientists. The result of conventional discovery processes are often not matching the domain context of the users and obligate them the tedious task of filtering large result sets to obtain the original object of the interest of the researcher intended to find with the search. Therefore the need arises for an improving discovery method, which will incorporate the domain knowledge and additional semantic information into the search in order to obtain a more fitting result for the specific context of the user.

The MU scenario not only aims to improve the discovery of scientific resources for one particular domain, but also tries to discover and create new relationships among different domains. The correlation of environmental pollutants including their transport due to global climate change and their health impact on the population is only one significant example of creating new relationships among different domains. These dependencies could represent new scientific insights for already available resources and connect the knowledge of the single domains. These relationships should facilitate further discovery process to deliver matching resources of multiple domains.

The MU scenario represents the close cooperation and joint venture of two university institutes: the *Research Centre for Toxic Compounds in the Environment*[2] (RECETOX) and *Institute of Biostatistics and Analyses*[3] (IBA).

RECETOX is an independent institute of the MU. RECETOX performs research, development, education and expertise in the field of environmental contamination by toxic compounds with specific focus on persistent organic pollutants (POPs), polar organic compounds, toxic metals and their species and natural toxins - cyanotoxins. It is also Stockholm Convention Regional centre[4] for capacity building and transfer of technology in Central and Eastern European countries. The *Stockholm Convention on*

[2] http://www.recetox.muni.cz
[3] http://www.iba.muni.cz
[4] http://www.recetox.muni.cz/index-en.php?pg=regional-pops-center

Persistent Organic Pollutants[5] (*Stockholm convention*) is a global treaty to protect human health and the environment from chemicals that remain intact in the environment for long periods, become widely distributed geographically and accumulate in the fatty tissue of humans and wildlife. RECETOX is formed by several research divisions, service laboratories and technology-transfer centres: Environmental chemistry and modelling, Ecotoxicology and risk assessment, Trace laboratory, and Laboratory of data analyses. Research and development of the centre include monitoring of environmental matrices, studies of environmental fate and effects (ecotoxicology) of toxic compounds, ecological and human risk assessment as well as the development of informational and expert systems.

RECETOX launched the first version of the *Global Environmental Assessment and Information System* (GENASIS)[6] in January 2010 [1]. GENASIS provides information support for implementation of the *Stockholm convention* at an international level [7]. The initial phase of the GENASIS project is focused on data from regular monitoring programmes of POPs, providing a general overview of spatial patterns and temporal trends of pollutants concentrations [10].

IBA is a research institute of the MU, which is focused on delivering solutions to research problems arising in the environment and human health and it is providing related services, especially in the field of biological and clinical data analysis, organization and management of clinical trials, software development and Information and Communication Technology (ICT) applications. IBA research activities are primarily focused on organizational and expert services for large scientific projects. IBA is formed by four divisions: Division of Data Analysis, Division of Clinical Trials, Division of Information and Communication Technologies, and Division of Environmental Informatics and Modelling. For example, IBA created the first web portal for epidemiology of malignant tumours in the Czech Republic, the *System for Visualizing of Oncological Data* (SVOD)[7], based on the data from the Czech National Oncology Registry [3].

Dušek [2] pointed in 2009:"*A full-area monitoring of the environmental risk factors in all main environmental components is performed in the Czech Republic. The main objective of this functional monitoring network is the estimation of exposure to xenobiotic substances, and the evaluation of subsequent risks to human health. The system provides information for health risks management and also serves for public education, which is a prerequisite for active care of one's own health. The outputs from monitoring systems may also be used for assessing human risks associated with cancer epidemiology. Data about POPs are of key importance, since these compounds are known to have a wide spectrum of carcinogenic effects, a tendency to bioaccumulation, and are subject to long-distance transport*."

The objective of the MU scenario is to use and validate the resulting tagging and discovery framework of the TaToo project. Since the primary scope of the TaToo project is to facilitate the discovery of environmental resources, this scenario delivers the perfect opportunity to validate the resulting solution against challenging real word problems. There are numerous scientific domains available and actively researched at

[5] http://www.pops.int
[6] http://www.genasis.cz
[7] http://www.svod.cz

the MU, but two important domains have been carefully chosen to demonstrate and validate the envisioned functionality of the TaToo project. The vision of the MU scenario is that other scientific domains could follow the initial institutes to further spin a new kind of knowledge network to deliver a new generation of tools and methods to effectively and conveniently support the scientific user in their daily work.

2 Align ontology for POP's and cancer domain

The MU case study ontology is aligned to the TaToo ontology framework through the TaToo Bridge ontology. Importing the bridge ontology enable to have all of the class, property and individual definitions from the Bridge ontology available for use in our proposed ontology. Figure 1 shows classes from the first version of the bridge ontology which are important in proposed align ontology.

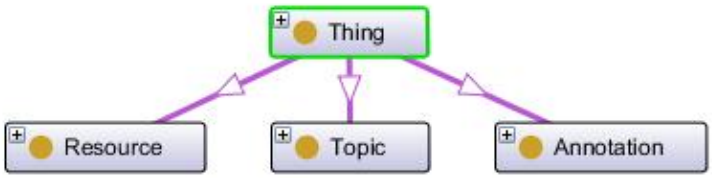

Fig 1. TaToo Bridge ontology

2.1 Analyze of POP's and cancer interests

By analyzing of both domains, i.e. environmental and human risk assessment areas we identify four levels of proposed ontology system.

The first one named **nomenclature** contains a set of nominal descriptors identifying key objects (variables) which should be mostly identified in investigation or in research searching for information (POPs – chemical compounds and diseases – cancer diagnoses). This level is highly standardized adopting internationally unified, extensively translated nomenclature systems. The proposed framework works with key (major) nomenclature system and additional (supporting) sub-systems. Proposed ontology architecture is relatively rigid in this dimension as the used nomenclature systems are based of widely accepted international consensus. In proposed version of ontology we used ICD-10 class hierarchy (International Classification of Diseases) and recommended POPs and Matrix taxonomy based on Stockholm Convention.

The second level is **classifiers** which present a attributes determining some key properties of the examined objects (chemical compounds, diseases). Only classifiers highly relevant for exposure environmental studies and risk assessment studies are adopted. In result, the classifiers represent binary codes or multiple categories, typically derived on the basis of some external information reachable from

standardized database, evidence-based literature, thesaurus or encyclopedia (properties of given chemical compounds, properties of disease at the time of diagnosis, etc.). The set of attributes is flexible according of used classifiers; e.g. In case of studies focusing on some special topics etc. This level is included in the proposed ontology.

Next level is **information source identifiers -** necessary descriptors of the source of information which is processed or needed. These attributes also refer to some type of validity scoring because they describe type of studies and other information sources which can be regarded as relevant. Furthermore, this set of attributes allows the users to specify studied problem or scientific field to be inspected. Representative from this level in proposed ontology is *ProjectType* showed in Fig. 2.

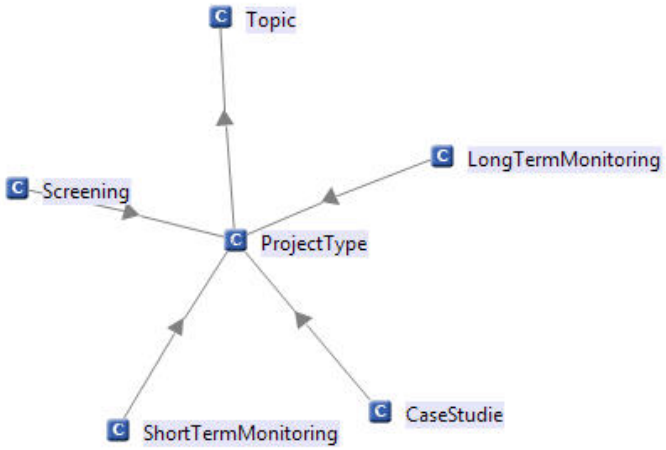

Fig 2. Example of information source identifiers - *ProjectType*

Last level contains **obligatory descriptors** which represent family of variables describing the lowest level of the information processed or searched for. It means set of variables which are obligatory to understand numerical values reached in searching engines (units, sampled, matrices, epidemiological measures, time and site specification, etc.). These descriptors are strictly obligatory and represent valid information, i.e. only information, numerical value supplied with these descriptors can be regarded as valid and trustworthy.

2.2 Identification of basic concepts

By analyzing of our two domains (POPs contamination and cancer epidemiology) we identified number of concepts. Below we describe several of them, which are necessary for understanding of proposed concept and proposed connections with MERM (Minimal Environmental Resource Model) concept [8]. The MERM concept

has been developed following the NeOn methodology [4]. The MERM describes the structure of resources and annotations in TaToo, being a reference data model for both services and user interfaces.

CancerTopic - defined as sub-concept of the MERM *Topic* concept and represent the possible keywords belonging into cancer risk domain.

POPsTopic - defined as sub-concept of the MERM *Topic* concept and represent the possible keywords belonging into POPs domain.

CancerAnnotation - defined as sub-concept of the MERM *Annotation* concept and represent the group of annotations connected with cancer risk domain.

POPsAnnotation - defined as sub-concept of the MERM *Annotation* concept and represent the group of annotations connected with POPs domain.

Matrix - define the taxonomy of possible matrixes and this concept is connected with *POPsTopic* sub-concept.

Disease - define the taxonomy of possible diseases. Because our domain is restricted to cancer epidemiology, we also define sub-concept C*ancer* and example class *BreastCancer*. These sub-concepts are also connected with ICD10 taxonomy of diseases. This taxonomy is used through connection with *Diagnose* class (sub class of *CancerTopic*).

In proposed ontology are also others concepts, but they have similar structure as classes mentioned above. We plan later extend these concepts with all aspects mentioned in chapter 2.1.

2.3 Alignment domain ontology to the TaToo ontology framework

The TaToo Ontology framework comprises the bridge ontology, a number of domain ontologies, and a number of alignment ontologies. The framework is not limited by a number of different domain ontologies, that is, by a number of different environmental sub-domains whose resources are managed by the TaToo system. However, in order to be able to plug-in a domain ontology in the TaToo ontology framework, the domain ontology needs to be accompanied by an appropriate mapping interface. This mapping interface is identified in the framework as an alignment ontology.

There are two possible alignment strategies for mapping these ontologies. The first mapping strategy, uses the *rdfs:subClassOf* construct defined in the *RDFS* to map concepts of the domain ontology to the concepts of the MERM and bridge ontology. The second mapping strategy, assumes the usage of the *owl:equivalentClass* and *owl:sameAs* constructs defined in OWL.

For mapping of our domains to the TaToo ontology framework we used the first strategy. Mapping is based on four ontology concepts: CancerAnnotat, POPsAnnotat, CancerTopic and POPsTopic. Fig. 3 illustrates these mappings and connections between them.

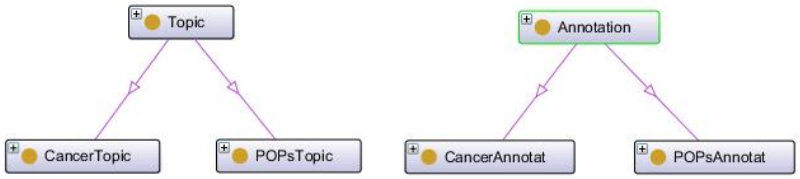

Fig 3. Mappings domain concepts to the TaToo ontology framework

2.4 Description of basic concepts

We describe in this section some of identified concepts in more detail. These examples are not trivial and show an interesting perspective of described domains from the ontology design point of view.

First one is illustrated at Fig. 4, where is presented *POPsTopic* class, which is the super class of *Topic* from the MERM ontology. It has sub classes *Matrix,* and *Compound* which are connected with POPs domain and enable to classify annotated resource. First contain information about used matrix and the second information about monitored compound.

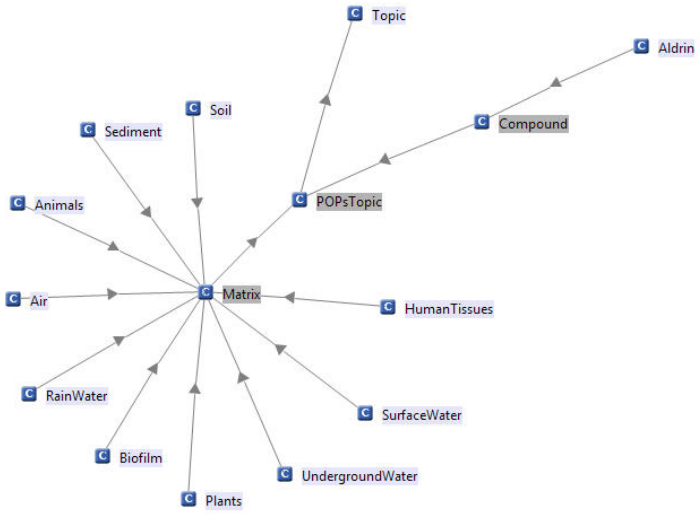

Fig 4. Definition of *POPsTopic* class

The next one describes the definition of *CancerTopic* class. We can see on Fig. 5 that this class has a super class *Diagnose* and a sub class *EpidemiologicalMeasures.* The class *Diagnose* is equal with the *Cancer* class which is sub class of *Diseases*. The

Diagnose class is equal with *C00_D48* sub class form ICD10 ontology (taxonomy of diseases).

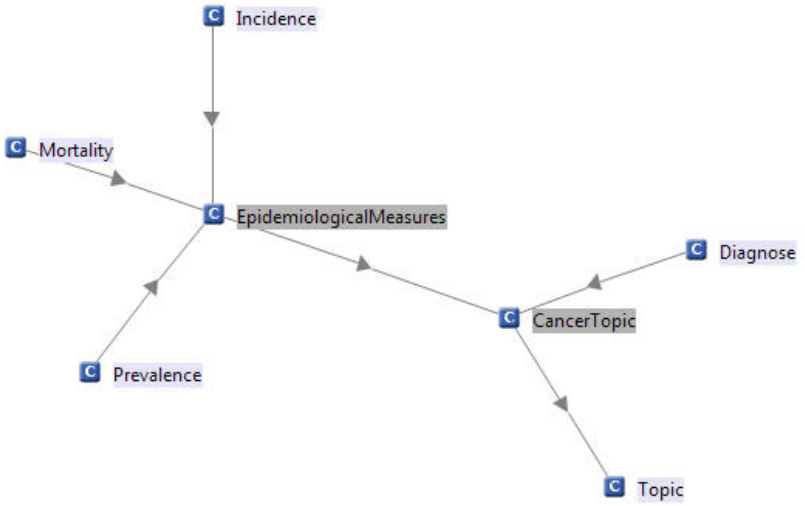

Fig 5. Definition of *CancerTopic* class

4 Conclusions

This paper contains the first initial proposal of align domain ontology for POPs contamination and cancer epidemiology domains. The development of domain ontologies in the TaToo Ontology framework is an evolutionary process, so we plan to extend them by adding new specific concepts and relationships that are currently not provided.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr. 247893.

References

1. Brabec, K., Jarkovský, J., Dušek, L., Kubásek, M., Hřebíček, J., Holoubek, I., Čupr, P., Klánová, J.: GENASIS: System for the Assessment of Environmental Contamination by Persistent Organic Pollutants. In: EnviroInfo 2009. Environmental Informatics and Industrial Environmental Protection: Concepts, Methods and Tools. 23. International

Conference on Informatics for Environmental Protection. pp. 369--376. Shaker Verlag, Aachen (2009)
2. Dušek, L., Mužík, J., Koptíková, J., Brabec, P., Žaloudík, J., Vyzula, R., Kubásek, M.: The national web portal for cancer epidemiology in the Czech Republic. In: Enviroinfo 2005. 19th International Conference Informatics for Environmental Protection. pp. 434--439. Masaryk University Press, Brno (2005)
3. Dušek, L.: Czech Cancer Care in Numbers 2008–2009, Grada Publishing, Praha (2009)
4. Gómez-Pérez A., Motta E., Suárez-Figueroa M.C., NeOn Methodology in a Nutshell. http://www.neon-project.org/nw/NeOn_Book (2010)
5. Hřebíček, J., Dušek, L., Kubásek, M., Jarkovský, J., Brabec, K., Holoubek, I., Kohút, L., Urbánek, J.: Anthropogenic Impact and Global Climate Change. Description of Validation Scenario in TaToo Project. In 7. letní škola aplikované informatiky. Indikátory účinnosti EMS podle odvětví. pp. 6-23, Brno : nakladatelství Littera (2010)
6. Hřebíček, J., Dušek, L., Kubásek, M., Jarkovský, J., Brabec, K., Holoubek, I., Kohút, L., Urbánek, J.: Validation Scenario for Anthropogenic Impact and Global Climate Change for Tatoo. In Proceedings of the Workshop "Environmental Information Systems and Services - Infrastructures and Platforms". CEUR-WS, Aachen (2010)
7. Klánová, J., Čupr, P., Holoubek, I., Borůvková, J., Přibylová, P., Kareš, R., Kohoutek, J., Dvorská, A., Komprda, J.: Towards the Global Monitoring of POPs - Contribution of the MONET Networks. RECETOX, Masaryk University, Brno (2009)
8. Pariente, T. et al.: A Model for Semantic Annotation of Environmental Resources: The TaToo Semantic Framework. In: ISESS 2011 (to appear)
9. Rizzoli, A., Schimak, G., Donatelli, M., Hřebíček, J., Avellino, G., Mon, J.: TaToo: tagging environmental resources on the web by semantic annotations. In: iEMSS 2010. International Congress on Environmental Modelling and Software. Modelling for Environment's Sake. pp. 1192--1199, iEMSS, Ottawa (2010)
10. Urbánek J, Brabec K, Dušek L, Holoubek I, Hřebíček J., Kubásek M. Monitoring and Assessment of Environmental Impact by Persistent Organic Pollutants. In: Diamantaras K, Duch W, Iliadis L, editors. Artificial Neural Networks -- ICANN 2010. Vol 6354. pp. 483-488, Springer, Heidelberg (2010)