

Validation Scenario for Anthropogenic Impact and Global Climate Change for TaToo

Jiri Hrebicek¹, Ladislav Dusek¹, Miroslav Kubasek¹, Jiri Jarkovsky¹,
Karel Brabec², Ivan Holoubek², Lukas Kohut² and Jaroslav Urbanek²

¹ Masaryk University, Institute Biostatistics and Analyses, Kamenice 126/3,
62500 Brno, Czech Republic
{hrebicek, dusek, kubasek, jarkovsky}@iba.muni.cz,

² Masaryk University, Research Centre for Toxic Compounds in the Environment,
Kamenice 126/3, 62500 Brno, Czech Republic
{kbrabec, holoubek, kohut, urbanek}@recetox.muni.cz,

Abstract. The synthesis of existing Persistent Organic Pollutants pollution monitoring databases with epidemiological data is considered for identifying some impacts of Persistent Organic Pollutants on human health. This task requires new rich data, services, and model 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 full and easy access to web resources. The use of TaToo tools together with the Global Environmental Assessment and Information System and the System for the Visualizing of Oncological Data is discussed as a TaToo validation scenario for anthropogenic impact and global climate change influence on the trajectory of Persistent Organic Pollutants.

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

1 Introduction

Persistent organic pollutants (POPs) [6] represent a long-term problem which is connected with the production, application, and disposal of many hazardous chemicals and their impacts on human health. The Research Centre for Toxic Compounds in the Environment (RECETOX) of Masaryk University (MU) is focused on research of the fate and biological effects of POPs and other toxic substances in the environment. RECETOX monitors these chemicals in air, soil, water and human milk, and supports the implementation of international conventions on chemical substances like the *Stockholm Convention on Persistent Organic Pollutants* [9]. The 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, accumulate in the fatty tissue of humans and wildlife, and have adverse effects on human health or on the environment. The Convention is

administered by the United Nations Environment Programme (UNEP) and is based in Geneva, Switzerland.

RECETOX closely cooperates with the Institute of Biostatistics and Analyses (IBA) of MU. IBA is a research institute oriented towards devising solutions for scientific projects and providing related services, especially in the field of environmental, biological and clinical data analysis. IBA created the *System for Visualizing Oncological Data* (SVOD) [10] – a web portal concerning the epidemiology of malignant tumours in the Czech Republic, which is based on data from the Czech National Oncology Register.

Specific effects of POPs can include cancer, allergies and hypersensitivity, damage to the central and peripheral nervous systems, reproductive disorders, and disruption of the immune system. Some POPs are also considered to be endocrine disrupters, which, by altering the hormonal system, can damage the reproductive and immune systems of exposed individuals as well as their offspring; they can also have developmental and carcinogenic effects.

In January 2010 RECETOX launched the first version of the *Global Environmental Assessment and Information System* (GENASIS) [1], [4] – a web portal which provides information support for implementation of the Stockholm Convention at an international level. The initial phase of the GENASIS is focused on data from regular POPs monitoring programmes, providing a general overview of spatial patterns and temporal trends of pollutant concentrations. The aim is now to try to establish whether there is a connection between the concentration of POPs and the occurrence of cancer in some regions.

This task requires new information and communication technology (ICT) tools for the discovery gap in environmental information resources which will be developed within the FP7 project TaToo (Tagging Tool based on a Semantic Discovery Framework) [8], [11] and share the vision of a *Single Information Space in Europe for the Environment* (SISE) [7]. It aims to develop tools allowing third parties to easily discover web resources (data, services and models) and to add valuable information to these resources. TaToo tools will be validated in three different validation scenarios [5]. MU is validating the TaToo tools in the *Validation scenario of anthropogenic impact and global climate change* (MU Scenario). It aims to improve the discovery of web resources in the domains of environmental pollution by POPs including influence of global climate change and epidemiology, and tries to find relationships between these domains.

2 Development of TaToo Tools for the MU Scenario

Current TaToo development at MU is outlined in this section. To achieve the goal of the MU Scenario, it is necessary to use data from national and international monitoring networks, and to discover and obtain as-complete-as-possible data sets representing anthropogenic impact. There is a great deal of various environmental data on the internet. For standard search engines (Google, Yahoo, etc) it is difficult to find (all of) them and choose the relevant data. Discovery, use, and reuse of relevant data require enhancements of meta-information descriptions, which can be achieved

through TaToo's semantically rich environment. In this context, MU intends to employ TaToo tools [5] and will validate their performance with respect to the tagging and discovery of semantically rich resources concerning anthropogenic impact and the influence of global climate change on the transport of pollutants.

The users of TaToo tools in the MU Scenario can be divided into three categories:

1. *Scientific users*: They are regular users with a scientific background and assumed ICT skills. They will use the system to discover resources from both domains (POPs, health issues). They will be able to find resources, find similar resources (having already found some items), compare resources, and also find connections between resources. Everything on a “read only” basis.
2. *Domain experts*: The group of domain experts includes users who have some additional functionality for scientific users. Domain experts can also evaluate resources and assign metadata to the resources. By means of such functions they will contribute to the information enrichment process.
3. *System administrators*: system administrators will be responsible for organisational and maintenance tasks in order to guarantee proper system functionality. This involves also user administration, system settings, problem solving, user support etc.

We can also categorize possible web resources for the MU Scenario as follows:

- *Primary resources*: Structured raw data e.g. cancer patient records (diagnosis, sex, age, etc.) or measurements such as time series of POPs (method, compound, substance etc.).
- *Secondary resources*: Aggregated or processed information based on primary data e.g. diagrams, analysis results, automatically generated reports, scientific publications, books etc. in the form of well known datatypes (PDF, doc, txt etc.)
- *Information services*: Internet based services which provide information from the first and second category. For example, Sensor Observation Services which provide Time Series for POPs in the form of compound measurement values (Web Services standard etc.).
- *Models*: Meta-information about mathematical and computational models for the calculation of POPs distribution in the environment or dispersion models, which are used to estimate or to predict concentrations of airborne pollutants emitted from sources such as industrial facilities, local heating or traffic.

The MU Scenario profits from the GENASIS and SVOD web portals (created at MU), which collect the desired data. We have focused on these two portals as representatives, but there will also be other resources from outside. For the validation of TaToo tools we propose in the MU Scenario eight use cases (UC), which we introduce briefly as follows:

UC1: *Discover resources with existing tools*. SVOD and GENASIS users will be provided with the possibility to indirectly use the TaToo functionality for discovering similar resources based on analysed objects in mentioned web portals.

UC2: *Generic discovery*. The goal of this use case is to deliver improvements regarding result relevance compared to conventional search engine results. The relevance of the resources to the search criteria should be improved so that the user receives more potentially interesting search results. This circumstance hopefully leads to a reduction in the tedious effort of scanning search results for matching entries.

UC3: *Persistent Organic Pollutant Resource Discovery.*

UC4: *Oncological resource discovery.* UC3 and UC4 bring extended domain specific search. Domain experts are allowed to enrich the resources by using TaToo tools.

This example UC will be introduced in detail in the next chapter.

UC5: *Define resource uncertainty.* Domain experts are allowed to define certain quality criteria for resources.

UC6: *Compare resources.* UC5 enables users to compare discovered resources immediately after the discovery. This use case should be helpful in finding the relationships between different resources either in the same domain or in different domains.

UC7: *Find similar resources.* UC7 allows the functionality to search for similar resources based on interesting resources already found.

UC8: *Find related resources.* The last use case provides the possibility to search for related resources in other knowledge domains based on an already found resource.

3 Oncological resource discovery - example of use case

This chapter introduces the proposed UC4 in more detail. The use case UC4 represents a domain specific search and has the goal of discovering resources with a focus on analysis and statistical results in the field of oncology. The user is interested in the discovery of cancer related resources based on a domain specific search mask with a common parameter such as diagnosis, gender, patient number etc.

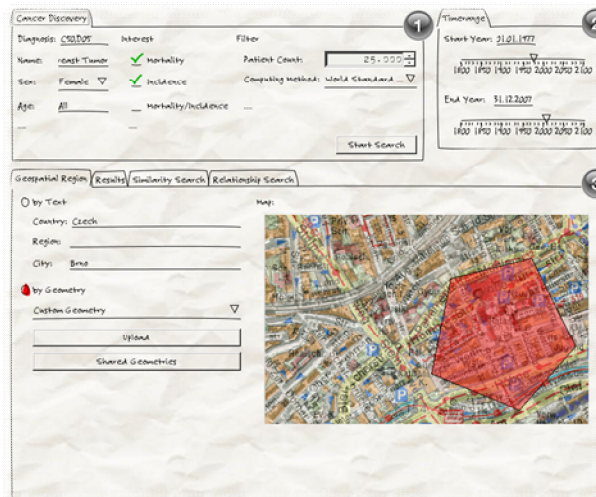


Figure 1. Cancer discovery mock-up

Figure 1 shows a possible mock up prototype for such a cancer search. The single panels will be explained in the following:

1. *Specify Search Parameter.* This panel illustrates what a specialised search mask for oncological discovery could look like. The user has the possibility to specify the

parameter for the search and set additional filters such as calculation methods, patient count.

2. *Specify Time Range.* The user has the option to define the start and end time of the period of interest. The result list should only contain information which has content within the temporal boundaries.
3. *Specify Geospatial Region.* The user will have the possibility to specify a geospatial region indicating the spatial dimension of the desired resource. Similar to the time range only resources will be found which address the specified area.

After the search is configured to the needs of the user the discovery of new resources can begin.

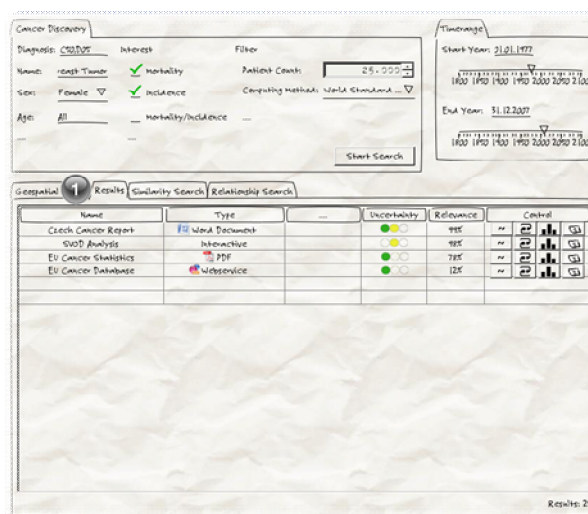


Figure 2. Presentation of results

Figure 2 shows a possible result list (indicated by one), containing discovered resources matching the search criteria entered by the user. For every resource, key values are displayed such as level of uncertainty, relevance, and additional actions. The user should be able to sort through these criteria, for example to find resources with a low uncertainty.

As a next step the user has the possibility to see a detailed description for every resource found. Figure 3 shows a possible visualization window of resources found in the previous search. The window is structured as follows:

1. *General Resource Information.* Information about the resource such as type and address of resource.
2. *Domain Specific Information.* Domain specific annotations of the resource are listed here. In this case, attributes such as number of participating patients, patient gender, and computing method are available and presented to the user.
3. *Related Resources* References to other resources. The related resources can be directly accessed and the description of the particular resource is presented to the user.
4. *Uncertainty Level.* Visualization of the uncertainty of the resource.

5. *Validation Information.* Meta information about the validation process, such as used tools, standards, and conclusion.
6. *Resource Visualisation.* A component to visualise contents of particular types of



resources.

Figure 3. Resource description

The above proposed forms will be built up with advanced web technology in TaToo [5], where the user can, for example, interactively select regions on a map, etc.

4 Conclusion

The above presented use cases will be implemented in the TaToo project in 2010 [5], [11]. Several of them can be generalized for all validation scenarios in the TaToo project since it is related to improving the discovery gap concerning environmental resources and to adding valuable information to the discovered resources. However, the MU Scenario is focused on two specific domains (POPs and cancer epidemiology) and aims at the discovery of connections between them. Discussed use cases UC6 - UC8 should help to find these relationships.

Acknowledgments. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement Number 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. Central and Eastern European Centre for Persistent Organic Pollutants, <http://www.recetox.muni.cz/ceepopsctr/index-en.php>
 3. 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)
 4. Global Environmental Assessment and Information System, <http://www.genasis.cz/main-index/en/>
 5. Lobo, T. P., Ciprian, M., Schimak, G., Avellino, G., Schlobinski, S.: Closing the discovery gap in environmental information resources using semantic annotations: the TaToo Approach. In: Enviroinfo 2010. 24th International Conference on Informatics for Environmental Protection. ENVIP 2010 workshop (to appear)
 6. Persistent organic pollutants, <http://chm.pops.int/Convention/ThePOPs/tabid/673/language/en-US/Default.aspx>
 7. Pillmann, W., Hřebíček, J.: Information Sources for a European Integrated Environmental Information Space. In: EnviroInfo 2009. Environmental Informatics and Industrial Environmental Protection: Concepts, Methods and Tools. 23. International Conference on Informatics for Environmental Protection. pp. 341-352. Shaker Verlag, Aachen (2009)
 8. 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. iEMSS, Ottawa (2010)
 9. Stockholm Convention on Persistent Organic Pollutants, <http://chm.pops.int/>
 10. System for Visualizing of Oncological Data, <http://www.svod.cz/?sec=aktuality&lang=en>
 11. Tagging Tool based on a Semantic Discovery Framework, <http://www.tatoo-fp7.eu/tatooweb/>