Seventh Framework Programme
ICT-2009-6.4
Information and Communication Technology

Tagging Tool based on a Semantic Discovery Framework

Project ID: 247893

Deliverable D5.1.9

Version 1.0

Refined Scenario Definition – Case 3 – V3

Due date of deliverable: 29/02/2012

Internal release date: 27/03/2012
Actual submission date: 03/05/2012

Document Control Page

Title	Refined Scenario Definition – Case 3 - V3	
Creator	MU	
Editor	Miroslav Kubásek, Jiri Hrebicek	
Description	This deliverable describes the main features of the Validation Scenarios for case 3, which focuses on the anthropogenic impact of global climate change	
Publisher	TaToo Consortium	
Contributors	MU, IDSIA, CIS	
Type	Text	
Format	MS Word	
Language	EN-GB	
Creation date	30.1.2012	
Version number	1.0	
Version date	3.5.2012	
Last modified by	MU	
Rights	Copyright “TaToo Consortium”. During the drafting process, access is generally limited to the TaToo Partners.	
Audience	☐ internal ☒ public ☐ restricted, access granted to:	
Review status	☐ Draft ☐ WP Leader accepted ☐ PCO quality controlled ☒ Co-ordinator accepted	Where applicable: ☐ Accepted by the GA ☐ Accepted by the GA as public document
Action requested	☐ to be revised by Partners involved in the preparation of the Project Deliverable ☐ to be revised by all TaToo Partners ☐ for approval of the WP Leader ☐ for approval of the PCO (Quality Manager) ☐ for approval of the Project Co-ordinator ☐ for approval of the General Assembly	
Requested deadline		

Table of contents

Index of figures

1. Management summary

1.1. Purpose of this document

This document contains the final version of the scenario definition - Case 3, which focuses on the evaluation on human health of the impact of anthropogenic causes of pollution combined with global climate change. The purpose of this document is to provide an overview of the scenario (background, objectives, and available tools), to define the types of users who will use the TaToo Tools, to describe the possible Use Cases of TaToo in the context of this scenario, and to provide an overall description of the functions of software applications to be developed.

We identified three types of users: scientific user, domain expert, or system administrator. In total, eight Use Cases are defined where some Use Cases are available only for certain types of users. Each Use Case is implements a specific functionality of the software applications.

The main resources the scenario is based on are the GENASIS and SVOD systems. The GENASIS system provides access to information on environment contamination by persistent organic pollutants (POPs). Combination of expert knowledge and validated data from several cooperating institutions create the opportunity for a broad spectrum of visualizations, analyses and modelling. GENASIS system is available on the website http://www.genasis.cz and it is offered in Czech and English languages. For more information about GENASIS system please see chapter 3.2.2. SVOD is a web portal about tumour epidemiology in the Czech Republic. It is primarily motivated by the effort to make the representative and valuable data of tumour epidemiology available to wide spectrum of users. Web portal SVOD is located on address http://www.svod.cz and is also available in Czech and English languages. For more information about SVOD portal please see chapter 3.2.1. At the end of the document, the definition of the metadata structure based on these two systems is given.

The objective of this scenario is to show how the TaToo Framework can be used to annotate, and then search and discover information stored in both GENASIS and SVOD, answering questions related to the impact of anthropogenic pollution due to POPs on human health. Validation scenarios will be created on the basis of the proposed Use Cases are described in more detail in chapter 4.

1.2. Intended audience

The targeted readers are Workpackage and Task Leaders of the TaToo project, in particular those of WP2, the Workpackage in charge of the analysis of user/communities needs, the existing technological gaps and the definition of the technical requirements in TaToo.

2. Abbreviations and acronyms

GENASIS	Global Environmental Assessment and Information System
CAS	CAS registry is the most authoritative collection of disclosed chemical substance information, containing more than 54 million organic and inorganic substances and 61 million sequences.
CMDSA	Cancer Minimal Data Source Attributes
CSU	Czech Statistical Office
ECB	European Chemicals Bureau
ECHA	European Chemicals Agency
ERD	Entity-relationship diagrams
IBA	Institute of Biostatistics and Analysis
ICT	Information and Communication Technology
IHCP	Institute for Health and Consumer Protection
IR	Information Retrieval
IUCLID	International Uniform Chemical Information Database
JRC	Joint Research Centre
MDR	Minimal Data Record
NIP	National Implementation Plan for the Implementation of the Stockholm Convention in the Czech Republic
NOR	National Oncological Registry
PMDSA	POPs Minimal Data Source Attributes
POP	Persistent Organic Pollutant
REACH	Registration, Evaluation, Authorisation and Restriction of Chemicals
RECETOX	Research Centre for Toxic Compounds in the Environment
SVOD	System for Visualisation of Oncological Data
UC	Use Case
UZIS CR	Institute of Health Information and Statistics, Czech Republic

2.1. Glossary

Anthropogenic
A term expressing: "caused by human activities"

Compound
Term "Compound" stands for "Chemical compound". Chemical compound is a chemical substance consisting of two or more different chemical elements. These elements consist of one type of atom. Common examples of elements are iron, mercury, lead etc.

Data source
A data source is any of the following types of sources for (mostly) digitized data: a database; a computer file; a data stream. Data from such sources is usually formatted and contains a certain amount of metadata.

Domain
Domain is a sphere of knowledge identified by a point of interest. Typically, in our context the knowledge is a collection of facts about some specialization (cancer, persistent pollutants, etc.).

Epidemiology
Epidemiology is the study of factors affecting the health and illness of populations, and serves as the foundation and logic of interventions made in the interest of public health and preventive medicine.

Fraction
A part of matrix (air, soil, water, etc.) such as different particle sizes (PM1, PM2.5, PM10), different parts (heights) of soil sample (0-2cm, 0-5cm, etc.).

Matrix
A part of animated or inanimate nature such as air, water, biota etc.

Mock-up
A Mock-up is a scale or full-size model of a design or device, used for teaching, demonstration, evaluating a design, promotion, and other purposes.

Oncology
Oncology is a branch of medicine that deals with tumours (cancer).

Persistent Organic Pollutant
Organic compound that persists in the environment for a long time (e.g. several years, or even decades of years), bioaccumulates, is toxic, and is subject to long-range transport.

Portal (http://looselycoupled.com/glossary/portal)
Web access point. A portal consists of web pages that act as a starting point for using the Web or web-based services.

Portlet (Java Portlet Specifications: JSR 168 and JSR 286, http://developers.sun.com/portalserver/reference/techart/jsr168/)
Portlets are web-based components managed by portlet containers that supply dynamic content. Portals employ portlets as pluggable user-interface components, a presentation layer, for information systems.

Primary Data
Data observed or collected directly from first-hand experience.

Prototype
A mock-up is called a prototype if it provides at least part of the functionality of a system and enables testing of a design.

Resource
A universal term for an information source like reports, scientific, time series, web services, web sites papers, publications, books, data sets

Resource Set
A set resources which are most likely related to each other, based on the semantically information of the different resources

Secondary Data
Published data and the data collected in the past or other parties is called secondary data.

Substance
Term "Substance" stands for "Chemical substance". Chemical substance is a material with a specific chemical composition.

Tagging
Adding Meta-Information to a resource.

TaToo Tool
A front-end component generally but not necessarily with a graphical user interface that allows or supports interaction with a human user. It resides on the Presentation Tier and acts as a client for the TaToo Public Services. Examples: the TaToo Portal, a client API library, a TaToo Toolbar.

Uncertainty
Uncertainty is the lack of certainty, a state of having limited knowledge where it is impossible to exactly describe existing state.

Use Case
Use Cases are the specific forms of using TaToo Tools by a certain Validation Scenario; see Validation Scenario.

Validation Scenario
Scenarios are used to validate the usability of TaToo's developments.

3. Description of the scenario

3.1. Background and objectives

This Validation Scenario, dealing with the correlation of environmental pollutants and their impact on population health, aims to show how TaToo can be exploited to create a central place for researchers, domain experts and decision makers to discover and access interdisciplinary knowledge in more efficient and usable way as what is currently expressed as 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. 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

originally object of interest the researcher intended to find. Therefore the need arises for an improved 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.

This Validation Scenario not only aims to improve the discovery of scientific resources for one particular domain, but also tries to discover and create new relationships between different domains. The correlation of environmental pollutants including their transport due to the effects of global climate change and their impact on the population health is only one significant example of creating new relationships between 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.

This Validation Scenario represents the close cooperation and joint venture of two university institutes: the Research Centre for Toxic Compounds in the Environment (RECETOX) and Institute of Biostatistics and Analyses (IBA).

RECETOX is an independent institute of the Masaryk University. RECETOX performs research, development, education and expertise in the field of environmental contamination by toxic compounds with specific focus on persistent organic pollutants (POP), polar organic compounds, toxic metals and their species and natural toxins - cyanotoxins. It is also a regional centre of the Stockholm Convention for Central and Eastern Europe. The Stockholm Convention1 on Persistent Organic Pollutants 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 development of informational and expert systems. In January 2010 RECETOX launched the first version of the Global Environmental Assessment and Information System (GENASIS)[2], which provides information support for implementation of the Stockholm convention at an international level. Initial phase of the GENASIS project is focused on data from regular monitoring programmes, providing a general overview of spatial patterns and temporal trends of pollutants concentrations.

IBA is a research institute focused on delivering solutions to questions arising in scientific projects and 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 applications. IBA activities are primarily focused on organizational and expert services for large scientific projects and clinical research 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

[1] http://www.pops.int
[2] http://www.genasis.cz

Modelling. IBA created a web portal for epidemiology of malignant tumours in the Czech Republic, the System for Visualizing of Oncological Data (SVOD)[3], based on the data from the National Oncology Registry (NOR)[4].

> *"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 persistent organic pollutants (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."* (Dušek, 2009)

The objective of the Validation 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 the Masaryk University, two important domains have been carefully chosen to demonstrate and validate the envisioned functionality of the TaToo project. The vision of the Validation 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.

3.2. Available tools and resources

3.2.1 System for Visualizing Oncological Data (SVOD)

Creating SVOD (System for Visualizing Oncological Data), a web portal about tumour epidemiology in the Czech Republic, was primarily motivated by the effort to make this representative and valuable data available to wide spectrum of users. Thanks to SVOD general epidemiology data about these serious diseases and related population risks is made freely available to everybody in the Czech Republic. Another aim of this web portal was to provide relevant information about tumour epidemiology in the Czech Republic abroad.

Web portal SVOD information services can be divided into three sections:

1) Current news: regularly updated information about population risk assessment and tumour epidemiology;

[3] http://www.svod.cz
[4] http://www.uzis.cz/registry-nzis/nor

2) Interactive analyses that allow the user to investigate directly epidemiological trends of selected oncological diagnoses;
3) Predefined presentations of important topics (Authorised information service).

These services are available freely to all users. All analyses contain only safe and publishable data of tumour epidemiology, without any personal data of patients.

The project of creating a web portal SVOD for tumour epidemiology in the Czech Republic is tied with longstanding development of software tools for the analysis of data coming from National Oncological Registry (NOR). The standalone software application SVOD (System for Visualizing of Oncological Data) was first created during the years 1999-2003 (now in version 6). This application made makes accessible all NOR data via a wide range of automated analyses. Although the software was finalized successfully, there were severe limits to its distribution and availability. The web portal SVOD solves all these problems and provides an effective way of access to epidemiological analyses to unlimited number of users (Figure 1).

The SVOD web portal works mainly with data from National Oncological Registry (NOR, 2007) which is managed by Institute of Health Information and Statistics (UZIS CR)[5]. It offers validated epidemiological data from the years 1977 - 2007. This represents a unique representative data set at least in European region (currently it stores 1 617 809 records). UZIS CR is therefore cited as a data manager in all outputs and is stated among scientific guarantees of the project.

Epidemiological trends cannot be made without relevant demographic data about examined population. This data was acquired from the Czech Statistical Office (CSU)[6] on the basis of general agreement about cooperation with Masaryk Memorial Cancer Institute[7] in Brno and Masaryk University (MU) in Brno.

The web portal SVOD was created by a team of authors from Faculty of Medicine MU in Brno (Centre of Biostatistics and Analyses) and Masaryk Memorial Cancer Institute in Brno. Creation of the portal SVOD is vitally supported by Ministry of Health of the Czech Republic in context of National healthcare quality programme. Further development is supported by a research programme of Masaryk Memorial Cancer Institute (Functional diagnostics of tumours, MZO 00209805) and a research programme of Faculty of Science MU (INCHEMBIOL - RECETOX project, No. 0021622412[8]). These grant projects guarantee long-term viability of the portal and ensure regular updates of data and successive development under supervision of administrators.

Information services of the web portal SVOD will be further developed, among others, on the basis of user suggestions and requirements. The main goal is to extend the information service in the area of population risk analyses in relation with available environmental data and other external risk factors (cooperation with above mentioned INCHEMBIOL project). The current version of the SVOD portal offers only epidemiological data, but NOR database allows

[5] http://www.uzis.cz
[6] http://www.czso.cz
[7] http://www.mou.cz/en/
[8] http://www.recetox.muni.cz/inchembiol/index.php?language=en&id=

even analysis of diagnostics and treatment of oncologic patients and survival analyses - even in relation to current hospital. These analyses are prepared for communication by the Oncological Society[9] and are available in the restricted zone of the portal SVOD. The web portal SVOD will therefore serve as an information source for Czech health management and can help to set up reference standards for healthcare results in oncology.

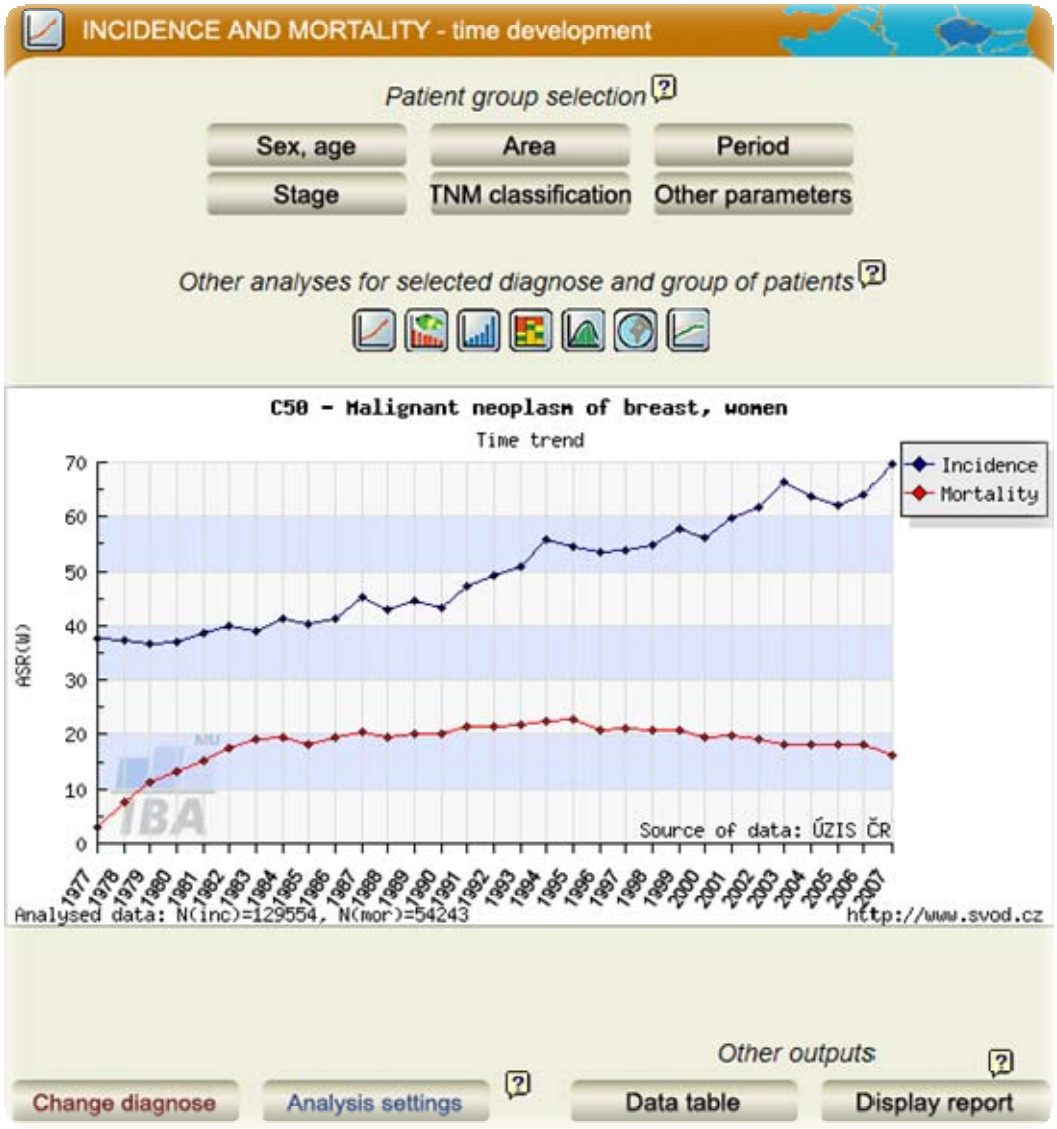

Figure 1. Breast cancer time trend - example of SVOD analyse

[9] http://www.linkos.cz/

3.2.2 Global Environmental Assessment Information System (GENASIS)

The GENASIS portal provides information support for the implementation of the Stockholm Convention on Persistent Organic Pollutants[10] at the international level. The information system is developed in accordance with the Single Information System of the Environment (JISŽP[11]) objective of the Ministry of Environment of the Czech Republic. Its connection with other data sources creates the potential for a comprehensive assessment of anthropogenic impact on the environment and the associated ecological and health risks. The portal GENASIS contains data collected by RECETOX and its partners since 1988 in various monitoring types (long-term, short-term, research studies, etc.).

The GENASIS portal also offers analytical tools, and this is one of the most important parts of the web portal. These tools allow basic processing of measured environmental data by using statistical program units. In the introductory screen the user can determine what kind of data will be analysed by the selection of various parameters (e.g. project name, sampling time, matrix, chemical compound, etc.). This procedure provides a core set of data. With tools implemented in this part of the GENASIS system it is possible to visualise the location of each sampled site by means of synoptic maps and examine general and / or detailed information about sampling frequency. It is also possible to sort and select / deselect localities and view measured concentrations of selected compounds at the localities.

However, mere visualization is not the main objective for the development of these powerful analytical tools. Using additional modules it is possible to obtain descriptive statistics for selected data set, observe changes in concentration of the user-selected chemicals during time period and easily depict seasonal and long-term trends (Figure 2).

Each module includes an option to use additional criteria that restrict entry data (e.g. selection of explicit altitudes). Another integral part of the analytical modules is stratification of localities according to various parameters (land use, altitude, distance to roads, sources of pollution, inhabited areas), which enables more detailed view and localities discrimination. More complex analyses and models are currently and continuously being prepared.

The pilot version of GENASIS project uses data from monitoring network MONET[12], which is focused on occurrence of persistent organic pollutants (POPs) in ambient air. But the primary motivation for GENASIS project is to make all representative and very valuable data about presence and distribution of these dangerous substances in the environment accessible for wide forum of users and interested public.

GENASIS project uses data collected both within the National Implementation Plan for the Implementation of the Stockholm Convention in the Czech Republic (NIP)[13] and international projects to reach its goals. The Czech Republic has had a long-term tradition in POPs monitoring in the environment and its monitoring networks cover all environmental components. A basic

[10] http://chm.pops.int/
[11] http://www.mzp.cz/cz/jednotny_informacni_system_zivotni_prostredi
[12] http://monet-cz.cz/index.php?pg=monet
[13] http://www.recetox.muni.cz/pops-centrum/index.php?pg=pops--nip

description supplemented by outputs used within the frame of the NIP is available for each monitoring network.

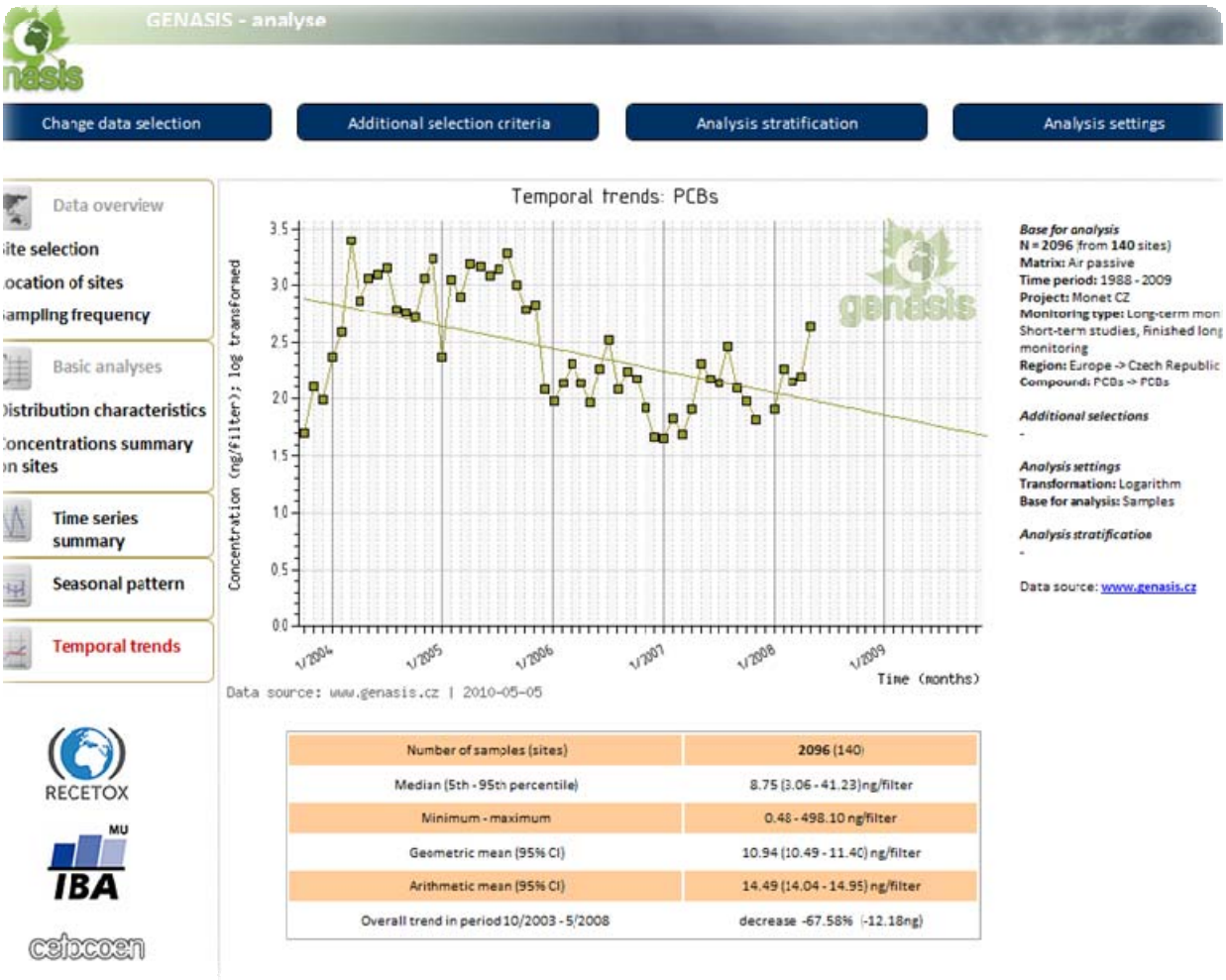

Figure 2. PCB compound time trend - example of GENASIS analyse

3.3. Relevance of the scenario

The synthesis of existing (air) pollution monitoring databases, with epidemiological data is required for identifying the effects of pollution on human health (anthropogenic impact). This task requires new, rich, data discovery capabilities within the bodies of available knowledge. IBA and RECETOX customers pose requests for new anthropogenic impact studies and influence of global climate change (e.g. a contamination of all environmental components by persistent organic pollutants through their changed transport due to global climate change) requiring data discovery from a multitude of monitoring networks and resources. Proper use of

such data requires contextual information, which TaToo will deliver through tagging and enhanced information description (meta-data) provided by an appropriate semantic environment. In this context, MU intends to employ TaToo Tools and validate their performance for tagging and semantic rich discovery of anthropogenic impact and global climate change resources.

TaToo developments will be demonstrated in the domain of anthropogenic impact and global climate change analysis. To achieve such a goal, it is necessary to use data from national and international monitoring networks, and to discover and obtain as-complete-as-possible data sets representing environmental anthropogenic impact. Discovery, use, and reuse of these data require enhancements of metainformation descriptions, which can be achieved through TaToo's semantic rich environment. In this context, MU intends to employ TaToo Tools and validate their performance for tagging and semantic rich discovery of resources of anthropogenic impact and the influence of global climate change on the transport of pollutants.

Climate is a factor strongly interacting with transport, transformation and effects of persistent organic pollutants in the environment. They are emitted into ambient air from various primary and secondary sources and atmosphere plays a key role in their transport both around their source surroundings and on long distances. Atmospheric transport is also main pathway of POPs transport into aquatic and terrestrial ecosystems. Current research of POPs global fate searches new information on sources, but also on other factors that affect pollutants concentration in ambient air, because climate processes at the interface of air and soil or water surface, and atmospheric transport significantly affect spatial and time variability of POPs in ambient air. From this point of view, regular measurements of pollutants concentration in ambient air at various localities and monitoring studies at various levels from immediate vicinity of local point sources up to continental level are of key importance. Important components of these measurements are monitoring design, selection of monitored chemicals, and selection of sampling and analytical methods, processing method, and data interpretation.

Scenarios of climate changes predict decreasing temperature contrast between poles and equator, drier continental interiors, wetter arctic and sub polar regions, modification of wind and precipitation patterns, sea level rise and others. All these environmental changes can influence the level of POPs in the environment, their partitioning among environmental compartments (air, soil, water), long-range transport, degradation rates, and toxic effects. Also, the release of POPs can be higher, for example due to pesticides usage to stop potential increase of malaria disease. Higher concentrations of POPs in the environment would then probably have more serious effects on living organisms.

The GENASIS system is based on database and linked analytical tools providing information base available on the web portal. Visualization of temporal and spatial patterns linked to characteristics of chemical compounds involved in Stockholm Convention support development of scenarios for individual environmental compartments.

Initial version of GENASIS system contains air pollution data and data from other matrices are being prepared. Analytical tools of GENASIS system provide visualisation of this data and basic statistics. The distribution models will be implemented in the near future to predict the fate of POPs in the environment. The user of TaToo Tools would be able to find and explore such

models from GENASIS website and also other relevant resources to investigate effects on the fate of POPs caused by global climate change.

TaToo Tools will be validated over specific scenarios and they will allow for continued collaborative development by federated users communities.

3.4. Overview of the scenario

A common task for the employees of the RECETOX and IBA, mainly research assistants and senior researchers, is the discovery of new information and resources in their particular research domain. Anthropogenic impact of global climate change scenario comes out from the joint research of the IBA and RECETOX. IBA has created a web portal of epidemiology of malignant tumours in the Czech Republic, SVOD[14]. RECETOX and IBA together develop a web portal called GENASIS[15], an expert information system based on POPs and Stockholm convention. With these two portals the natural question arises: Is there any relation between cancer occurrence and environmental concentrations of POPs?

To answer this question we must have relevant and appropriate (time and space) records for the incidence of cancer and for the measured concentrations of POPs. For this reason the minimum data standard was brought up both for SVOD and GENASIS. This minimum data standard explicitly identifies each record and is always required.

3.5. Resource categories

Individual types of information sources are linked to different ways of discovery. Raw data can be visualized in charts, queried or aggregated on flexible user demands. They are associated with detail information, easy data management but usually spatial or temporal constraints and limited access as well. More general are metadata allowing metadata analyses. The most available are fragmented information published on the web or printed in form of charts, summary tables, reports, inventories, encyclopaedias and information systems. Such variable information can be explored by review approaches. For this scenario we categorized possible resources as follow:

1. Primary information - Structured raw data e.g. cancer patient records (diagnosis, sex, age, etc.) or measurements like time series of persistent organic pollutants (method, compound, substance etc.).

2. Secondary information - Aggregated or processed information based on primary data e.g. diagrams, analysis results, automatically generated reports, scientific publications, books, etc. in form of well known data types (PDF, doc, txt, etc.)

3. Information services - Web Services (WS) which provide information from the first and second category. For example Sensor Observation Services which provide Time Series for persistent organic pollutants in the form of compound measurement values.

[14] http://www.svod.cz
[15] http://www.genasis.cz

3.6. Type of users

This chapter introduces shortly the different types of users. The users are divided into three categories:

1. **Scientific users**: scientific users are regular users with scientific background and assumed IT skills. They will use the system to discover resources from both domains (POPs, health issues). They will be able to find resources, find similar ones (having already found some resource), compare the resources, and also to find connections between resources. Everything on the "read only" basis.

2. **Domain experts**: group of domain experts collects users who have some additional functionality to scientific users. Domain experts can also evaluate resources and assign metadata to the resources. By the means of mentioned 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.

4. Use cases

The purpose of this chapter is to introduce several Use Cases. For our scenario we propose eight Use Cases that will be used for the validation of TaToo Tools.

For each Use Case we provide a description, a beneficiary and the expected benefit, the user, and the improvement indicators that will allow measuring the impact of the TaToo Tools.

4.1. UC1 – Discover resources with existing tools

This Use Case provides the users of SVOD and GENASIS with the possibility to indirectly use the TaToo functionality for the discovery of similar resources based on analysed objects. The TaToo discovery could be started from within the web analysis tools. The relevant information needed for the search would be already entered via the web interface during the analysis and can be submitted to the TaToo framework. The following paragraph shows a simple workflow for a TaToo enabled SVOD application.

Let us assume the user does not only want time trends of the Czech Republic, but also to discover information about other countries. Therefore the already selected domain information from SVOD such as: diagnosis - C50, D05, gender - female, cancer name - breast cancer etc. could be used to trigger a TaToo discovery. Figure 3 shows breast cancer incidence and mortality rate from 1977 to 2007 for the Czech Republic generated by the SVOD application. Number one indicates a simplified TaToo button, which the user would press to search for similar breast cancer related resources. UC4 illustrates how the TaToo discovery could be

preconfigured by a call from SVOD. Comparing UC1 and UC4 figures it is possible to notice that the domain specific information and general attributes like the time range are the same. The user only has to choose the desired geospatial region and to add additional attributes, which are not considered in SVOD to start the TaToo discovery.

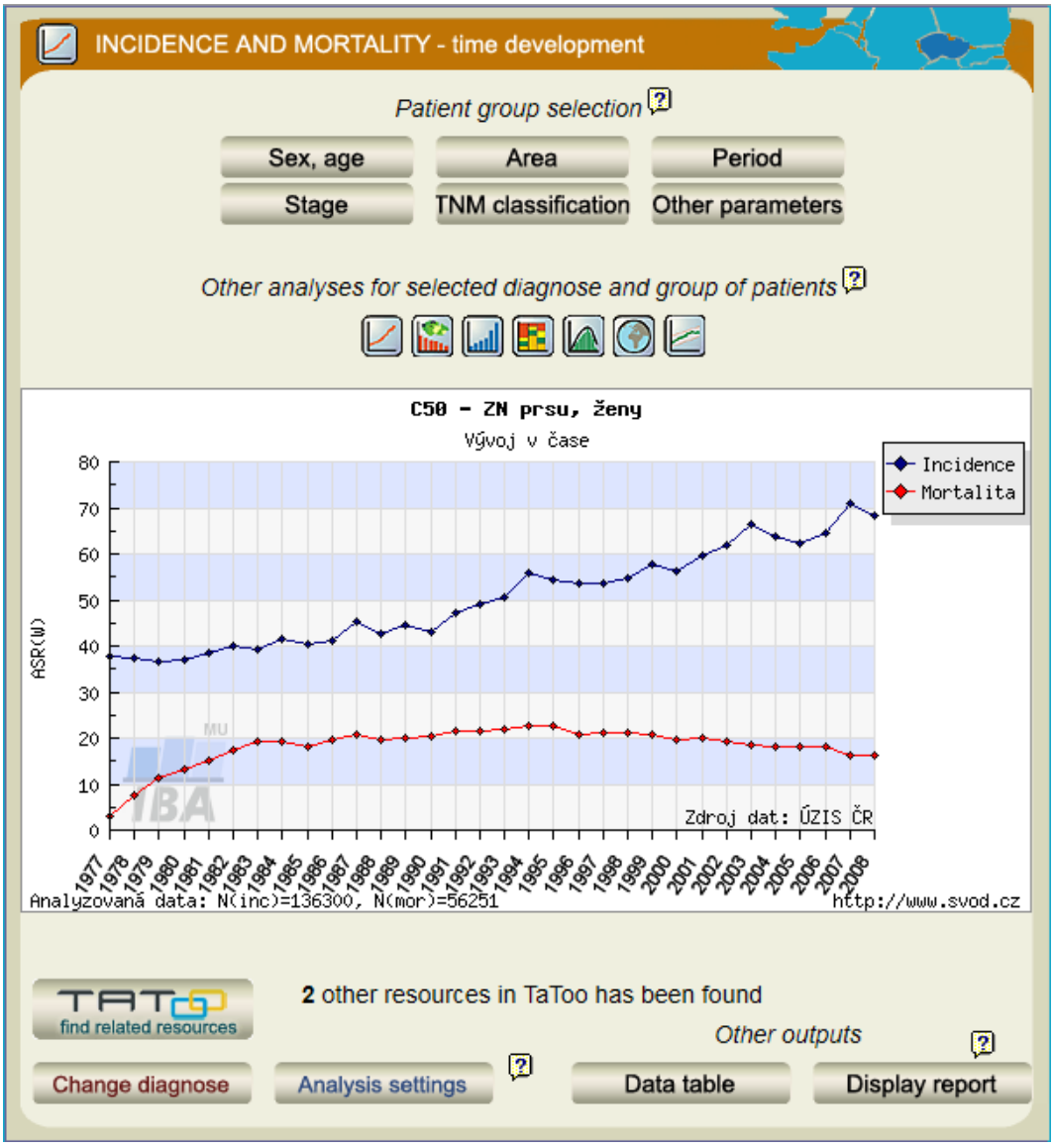

Figure 3. Integration of TaToo functionality in SVOD portal (live demo)

A simple example for SVOD integration could be to find similar cancer trend statistics for a particular diagnosis. The interface to TaToo will enable the user to search not only in records for the Czech Republic but will also provide the opportunity to discover similar time-trend analysis for other countries. The Resource Consumer will benefit from the fact that the currently used tool will formulate the query based on the analysed object and present the results delivered

by the TaToo framework. The query will be formulated on the basis of the already made choice of the user (e.g.: selection of region, diagnosis, gender of patients).

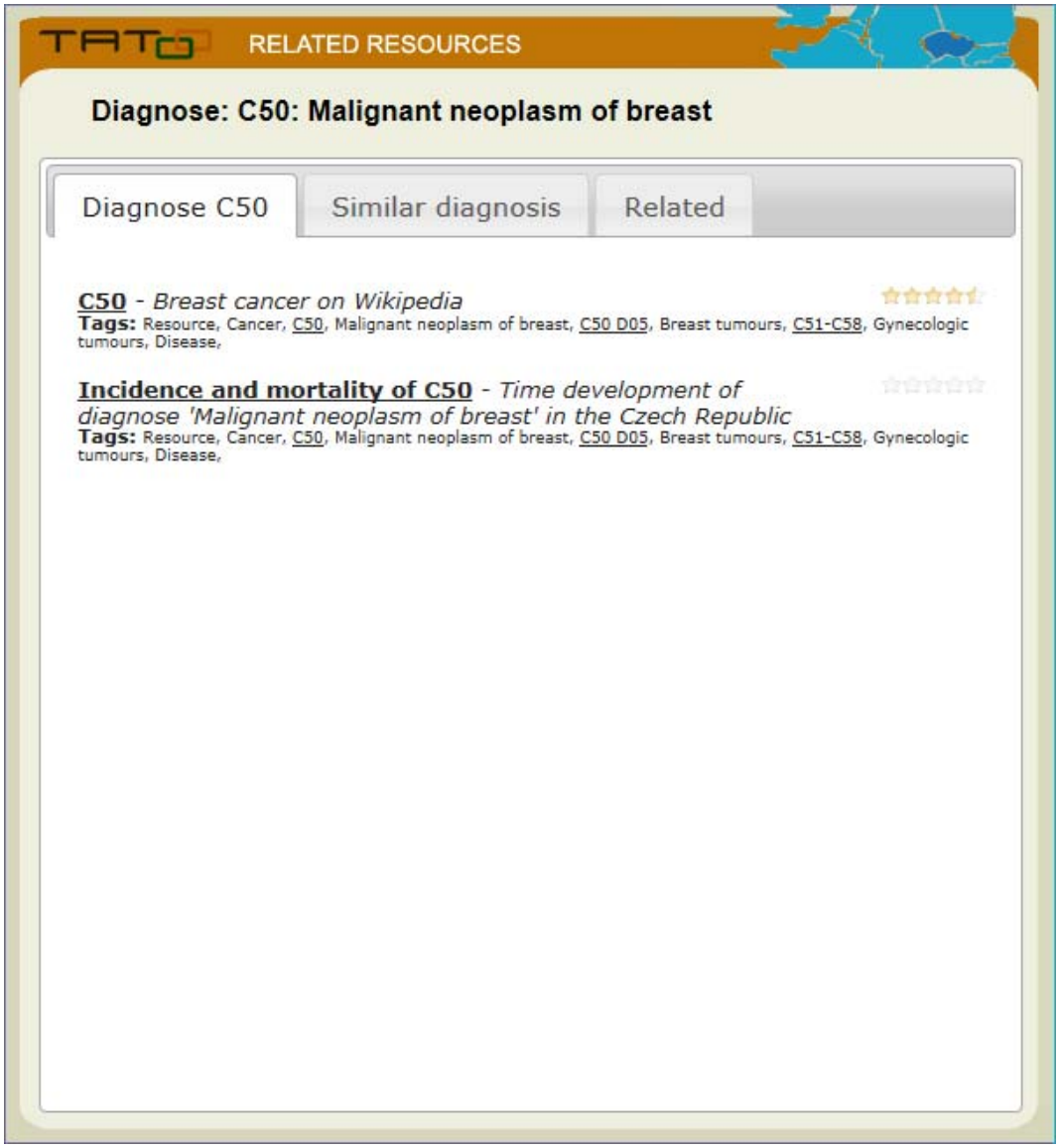

Figure 4.Extension of SVOD portal with TaToo results (live demo)

Well approved evaluation methods form the field of Information Retrieval (IR) could be used as improved indicators for the discovery process. Indicators such as:

- Recall - Fraction of the relevant documents which has been retrieved.
- Relative Recall - Ratio between the recall a number of relevant documents found.

- Recall Effort - Ratio between the relative recall and the number of documents examined to find the expected relevant document.
- Precision - Fraction of the retrieved documents which is relevant.

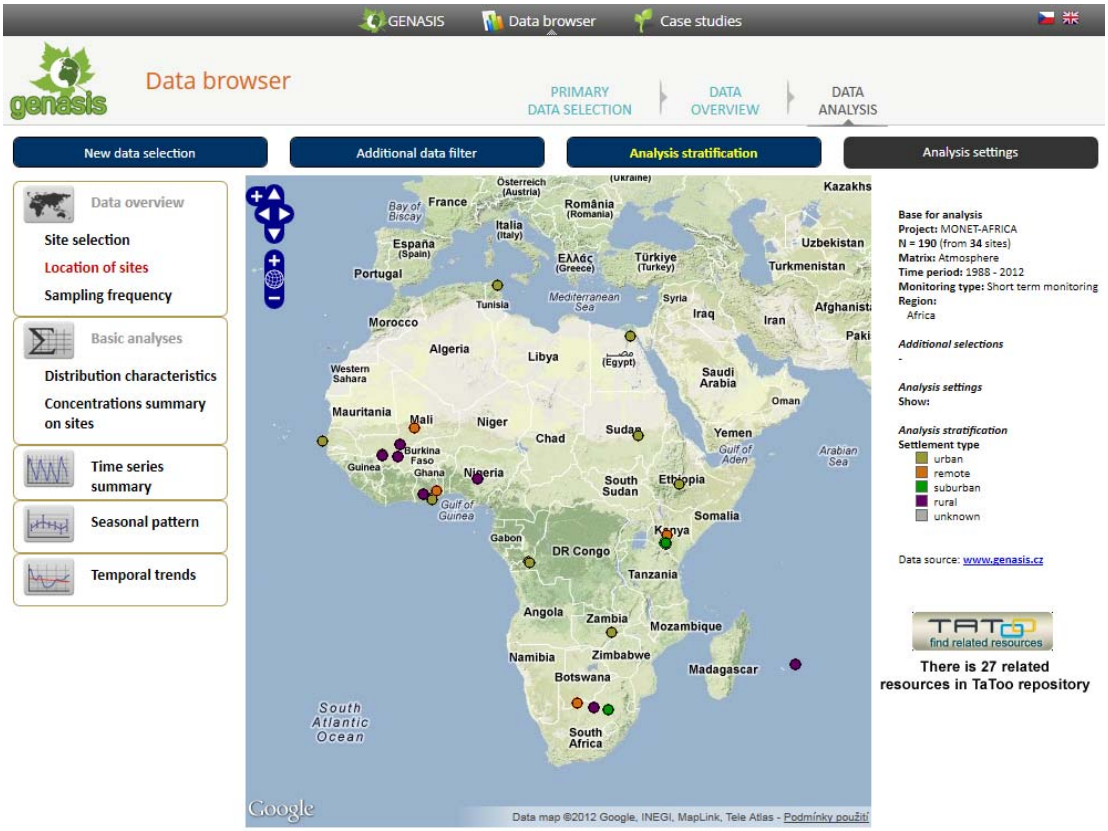

Figure 5. Integration of TaToo functionality in GENASIS portal

- Coverage - Fraction of the documents known to the user to be relevant which has actually been retrieved.
- Novelty - Fraction of the relevant documents retrieved which was unknown to the user.
- Expected Search Length - Average number of documents that must be examined to retrieve a given number of relevant documents.
- Satisfaction - Based on the Expected Search Length takes only the relevant documents into account.
- Frustration - Based on the Expected Search Length takes only the non-relevant documents into account.

The above list is not exhaustive and only serves as an overview of possible quality measurement indices, for further Information please see Baeza-Yates, 1999. Some of the above

mentioned measurements from the IR field could be adopted to validate the improvement of resulting TaToo functionality.

This Use Case has been implemented and is available as a live demo for validation of usability by experts.

4.2. UC2 – Generic discovery

In this Use Case the user wants to discover resources of a particular domain of interest matching certain criteria and keywords. The goal of this Use Case is to deliver improvements regarding result relevance compared to conventional search engine results. The user wants to find from the multitude of available resources the most interesting for his particular case, which is represented by the entered information. The found result should therefore have a higher probability to fit the desired domain context. Additional to domain specific information the system should also include other dimensions in the discovery like time range and geospatial information of the results, in order to further specify the domain of interest. The resulting list of resources should include additional information to the resource such as relevance to the search query, uncertainty information about the quality of the resource contains, file type of the resource etc. The search should not only deliver results in the original language used to specify the search query, it should also deliver results in foreign languages which match the domain context (e.g. user type search query in Czech language and TaToo will be able to understand and give to the user also English resources fulfilling typed query). The following two Use Cases, UC3 and UC4 represent a specialisation of this Use Case.

The benefit for providers of resources is that they get a better location of their offered resources. Resource consumers will get a new quality level of searching resources. The relevance of the resources to the search criteria should be improved so that the user receives more potential interesting search results. The amount of work needed to browse the results of conventional non semantic enhanced search engines will decrease because of the improved relevance of the search results. At least it will be also possible to discover resources which would be not discovered otherwise for example because of the barrier of language.

Initial version of this Use Case has been implemented as a part of TaToo Portal and is available for testing purposes.

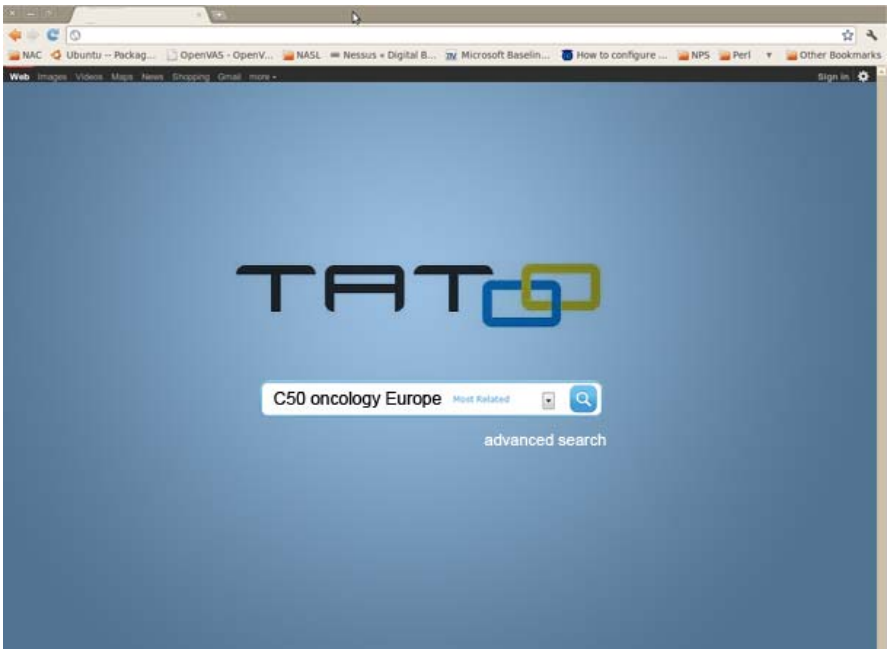

Figure 6. Proposed interface for TaToo generic discovery

4.3. UC3 – Persistent Organic Pollutant resource discovery

This Use Case enable users to find information in the domain of bio chemistry, specifically about Persistent Organic Pollutants (POP) monitoring. The interesting resources range from raw data such as time series with the actual measurements and additional information about measurement methods; measured compounds etc. to high level information generated from this raw data such as statistics and time trends of pollutants (Figure 7).

Initial version of this Use Case has been implemented as a part of TaToo Portal and is available for testing purposes.

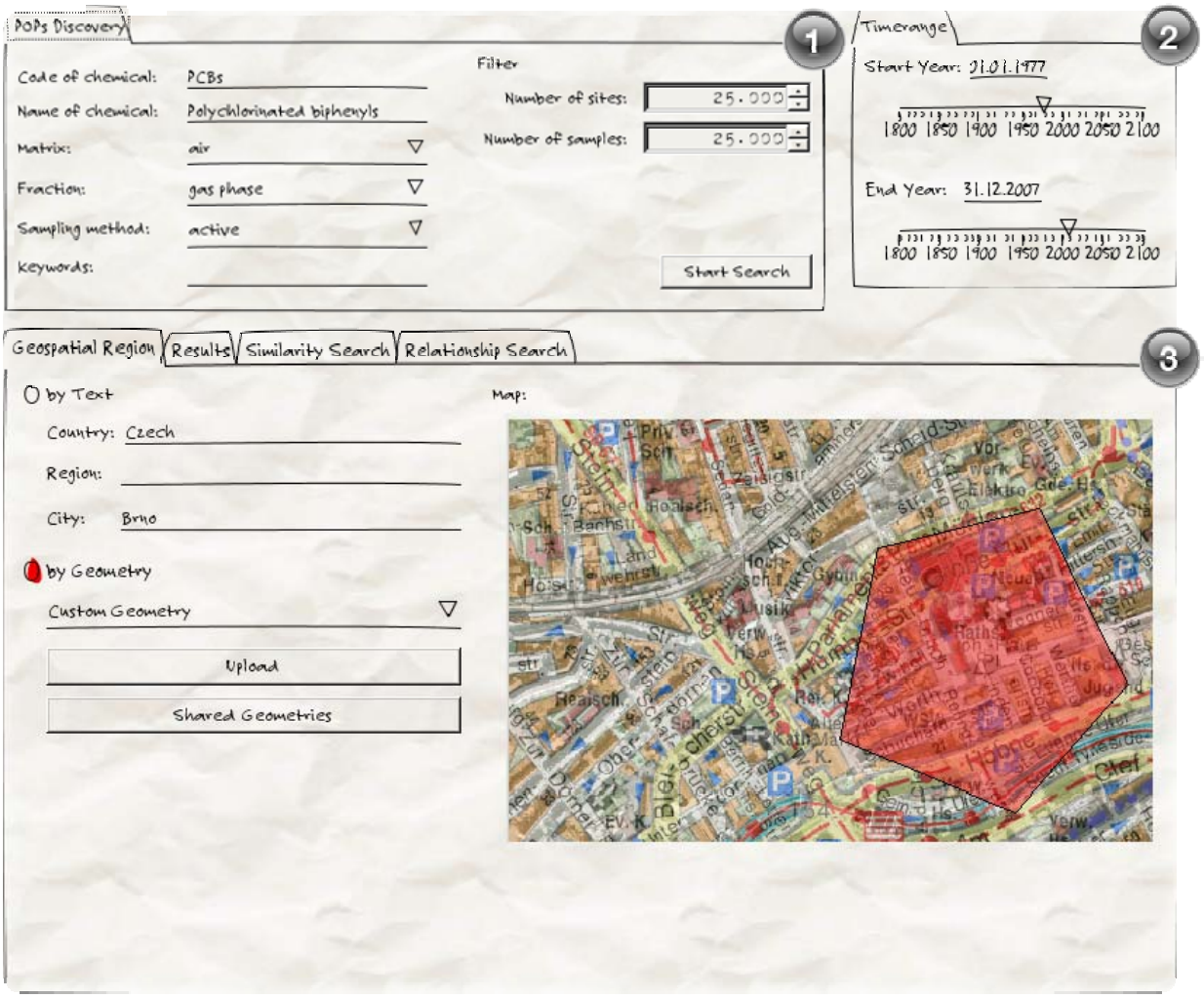

Figure 7. POP discovery mock-up.

4.4. UC4 – Oncological resource discovery

Similar to UC3 this Use Case represents a domain specific search and has the goal to discover resources with the focus on analysis and statistical results in the field of oncology. The user is interested in the discovery of cancer related resources, but unlike in UC3 it is most likely that there will be no patient records or raw data such as patient records because of confidentiality and data privacy policies. Nonetheless researchers are interested in discovering resources containing evaluations statistics, and reports regarding cancer incidences and mortality rates. Similar to UC3 the user wants to discover resources based on a domain specific search mask with common parameter such as diagnosis, gender, patient number, etc. In Figure 8 we propose of GUI for such a cancer search.

Initial version of this Use Case has been implemented as a part of TaToo Portal and is available for testing purposes.

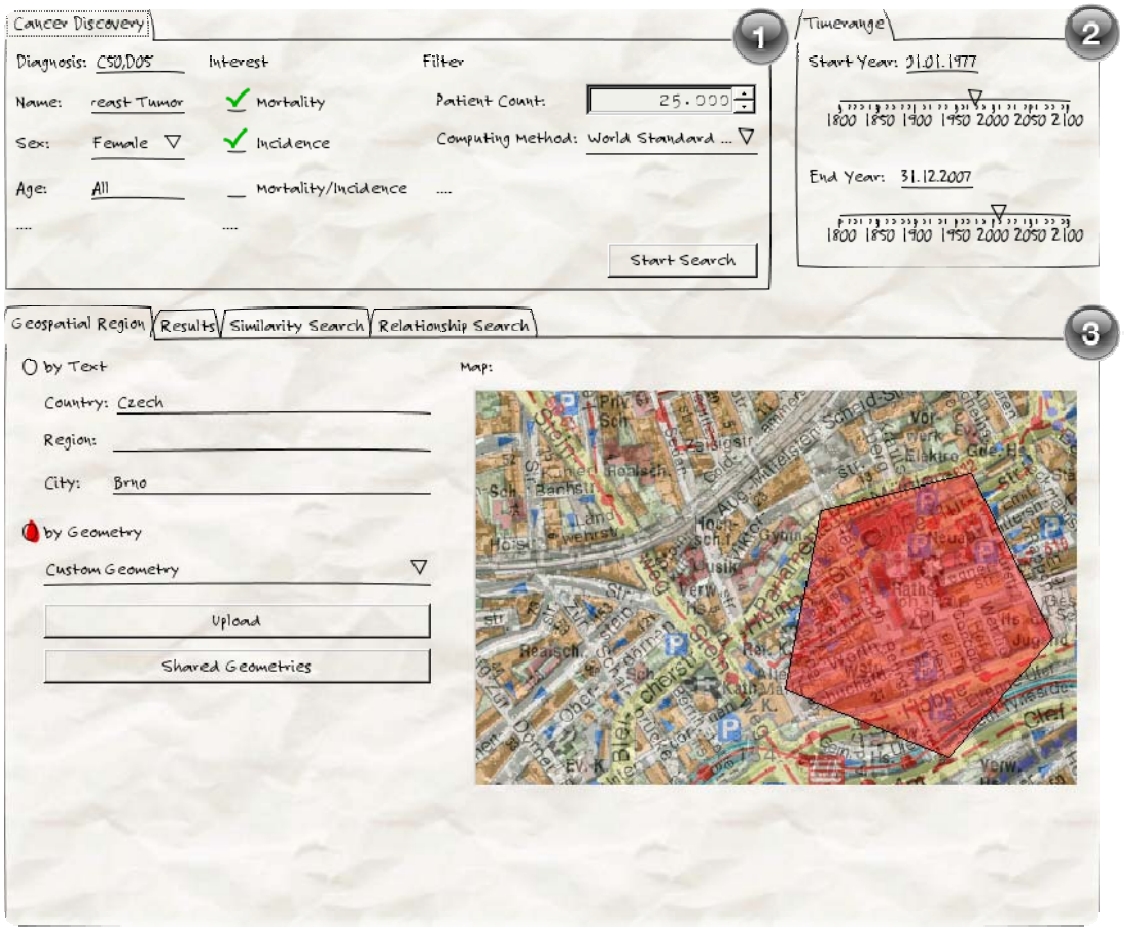

Figure 8. Cancer discovery mock-up

4.5. UC5 – Define discovered resource uncertainty

This Use Case should allow domain experts to define certain quality criteria for resources like the reputation of the publishing institute, the measurement methods, used norms and standards etc. The user should have the possibility to assess the different criteria with a value. Based on the different weighted criteria an uncertainty propagation level will be calculated and visualised in graphical and numerical way. However due to the fact that the optimal set of uncertainty criteria are not known, further investigation on this field has to be done. Therefore in this document only a fixed set of criteria will be used to calculate the uncertainty levels. Later version will further improve with advancing knowledge on how to represent the uncertainty of resources in more optimal ways. Figure 9 shows some of the initial criteria for version one.

Figure 9. Proposed uncertainty level detail mock-up

Proposed Uncertainty Criteria

Criteria are dependent to the domain of the data source and are computed based on metadata information. As initial approach we propose the following uncertainty criteria:

Cancer:

- Number of patients -1000 bad, 1000-5000 good-bad, 5000+ good - *this is a proposed dividing, it can be changed by the administrator or it can be also specified to the e.g. diagnose (different diagnose has various frequency)*
- Time range - 1 year bad, 2-5 years good-bad, 5+ years good - *this is a proposed dividing, for duration longer than 5 years we can see the trend, shorter durations are not usable in statistics calculations.*
- Size of region - has to be compared to the Number of patients, compute "Number of patients"/"Size of region" e.g. - *this rate describe how dense is monitoring. Bigger number is better.*
- Data source owner - True/False - *owner is known or not (binary criterion)*
- Data source provider - True/False - *provider is known or not (binary criterion)*

POPs:

- Number of sites- 3 bad, 4-10 good-bad, 10+ good - *this is a proposed dividing, it can be changed by the administrator.*

- Number of samples - has to be compared to the number of sites, compute "Number of samples"/"Number of sites" e.g. - *this rate describe how many samples has been average collected from one site. Bigger number is better.*

- Analysis method - True/False - *method is known or not (binary criterion)*

- Time range - 1 year bad, 2-5 years good-bad, 5+ years good- *this is a proposed dividing, for duration longer than 5 years we can see the trend, shorter durations are not usable in statistics calculations.*

- Size of region - has to be compared to the Number of sites, compute "Number of sites"/"Size of region" e.g. - *this rate describe how dense is monitoring. Bigger number is better.*

- Data source owner - True/False - *owner is known or not (binary criterion)*

- Data source provider - True/False - *provider is known or not (binary criterion)*

The resource provider would benefit from the uncertainty evaluation of the offered resources by domain experts. The provider will be able to see how external scientists judge the uncertainty of his data. The resource consumer will gain two advantages through this feature. First if the user is a scientific expert he will be able to provide feedback to the uncertainty level of resource. Second the user gets a first impression based on the uncertainty levels of the discovered resources how variable the data sources are and thus a quality indicator for the found resource. The discovery performance will be improved if an uncertainty evaluation for a resource is available, due to the fact that other users don't have to judge the quality of the resource on their own. This means if a resource is already analysed and a sound uncertainty level is determined it will facilitate further searches for qualitative resources. Discovery process will be presented for users with a new kind of feature, the inclusion of an uncertainty value as a search parameter. This enables searches for resources with a certain quality level.

4.6. UC6 – Compare discovered resources

This Use Case intends to enable a user to compare found resources on the fly after the discovery. Possible resources are for example PDF and office documents, different image formats, or raw data sets. The user should be able to add found resources to a compare list; this list serves as a temporal storage to enable the user to do multiple searches and to mark the potential resources for later comparison. The compare component itself should offer similar functionality as shown in Figure 10.

Number one in the prototype indicates an overview of the marked resources from previous performed searches. For every resource the recognised representations should be listed, for example found images within the content.

Number two visualises the actual comparison section, the user wants to add different resource representation here from the overview list. In the example prototype two time series

graphs are selected from the overview and are displayed to the user. One is a cancer trend from the SVOD application and one is a pollutant measurement from GENASIS, both diagrams can be compared by the user. Feasibility and practical interest for this feature has to be researched further.

The user will benefit from the possibility to actually compare well known parts of resources on the fly and thus is able to work more efficient and convenient.

The feature described by this Use Case, can be used as a preview function and serves as an information filter, due to the fact that only selected content of the resource is offered to the user, which implies antecedent scanning of the resource. This means that the user, if the data format is known to the system, does not necessarily have to browse the resource itself which should lead in some cases to an improved performance in discovering and comparing resources

The previous paragraphs already suggest and explain that through this feature a new kind of discovery process is offered. Through the scanning of the actual resource content, where applicable, selected information is accessible to the user without manual search in the resource.

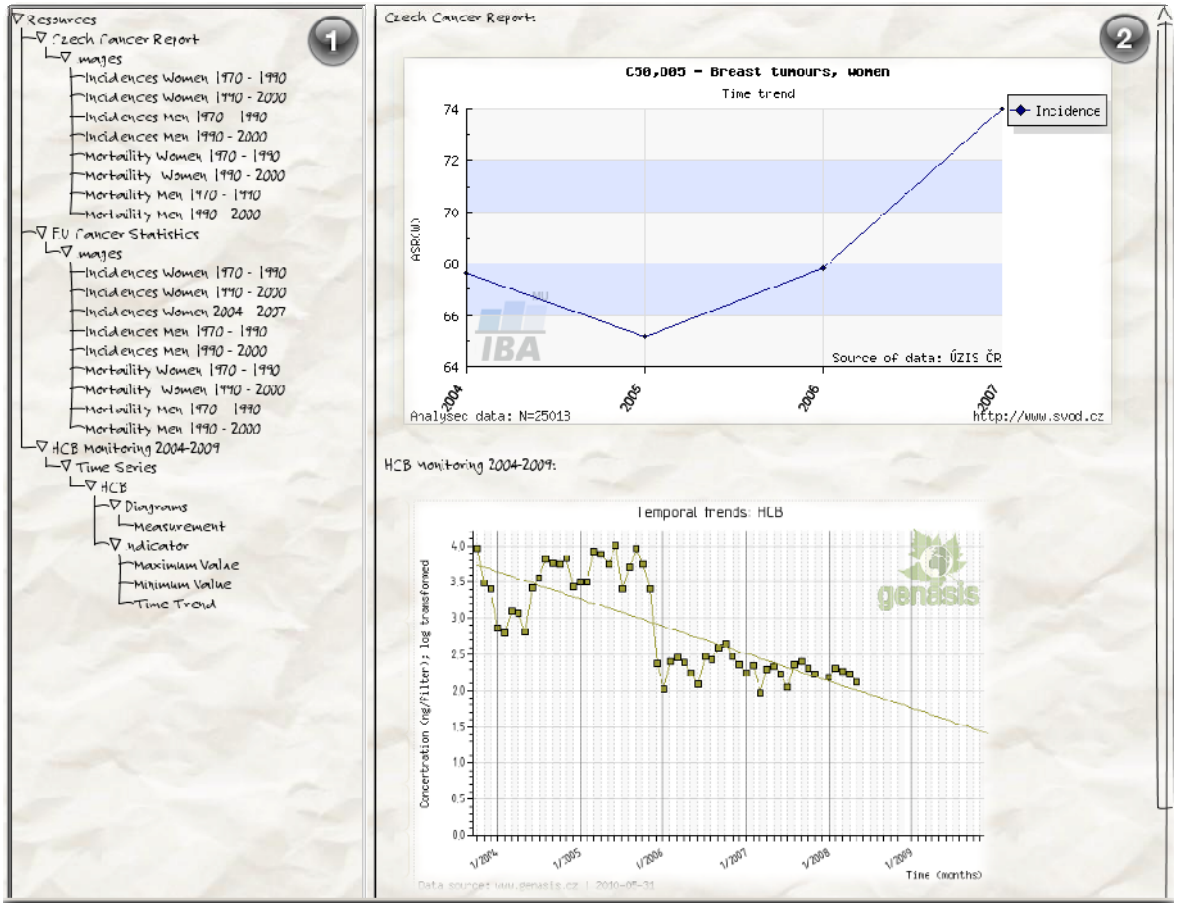

Figure 10. Compare resources tool mock up

4.7. UC7 – Find similar resources

This Use Case provides the functionality to search for similar resources based on an interesting resource already found. If the user finds a resource that matches his needs a new search can be initiated based on a current resource. The found resources should have a high probability to match the template resource used for the search.

Benefit of this Use Case is for resource provider that provided resources similar to other resources can be discovered more easily and accurate by the user. Resource consumer benefits from the feature to find similar resources to a template resource. The template resource metainformation will be used to further refine the initial search criteria; this will facilitate the search for the user. This Use Case will improve discovery process, because the user does not have to modify the initial search according to found resources, instead a new search can be started directly.

This Use Case has been implemented as a part of UC1 – Discover resources with existing tools and is available as a live demo for validation of usability by experts.

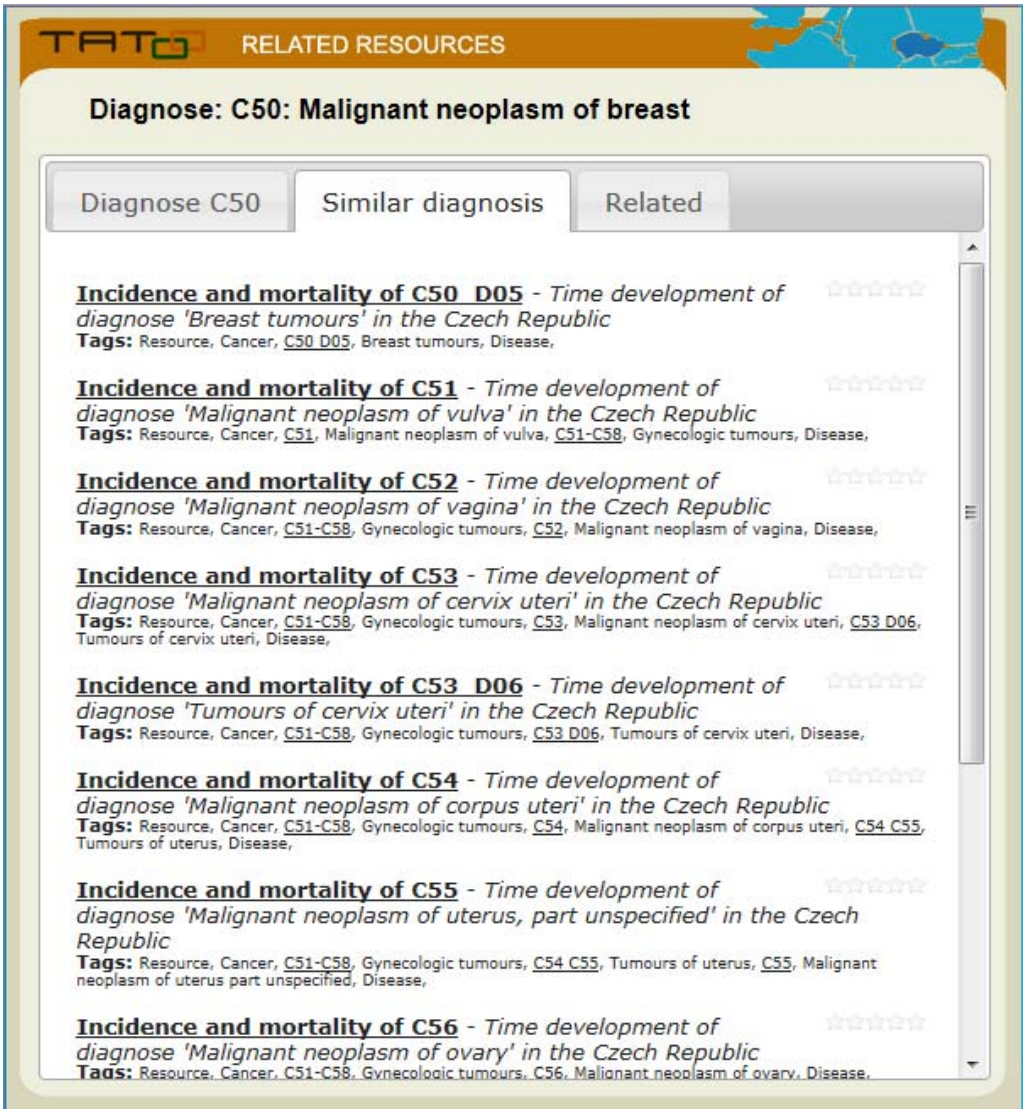

Figure 11. List of similar resources in SVOD portal (live demo)

4.8. UC8 – Find related resources

This Use Case covers the following that a user wants to search for related resources in other knowledge domains based on an already discovered resource. For example the user wants to find pollutant monitoring data for a specific time range and geospatial region, based on the values of a discovered cancer analysis. The geospatial extend and temporal extend from the cancer analysis will be used to perform a new search. The user only has to provide and specify the domain of interest in which new resources should be discovered.

User benefits from the feature to find related resources to a resource from other knowledge domains. The current resource metainformation will be used to find related resources in other domains. The discovery will be improved, because the user does not have to create a complete

new search, it is possible to reuse parts of the current search such as time range and geospatial region.

This Use Case has been implemented as a part of UC1 – Discover resources with existing tools and is available as a live demo for validation of usability by experts.

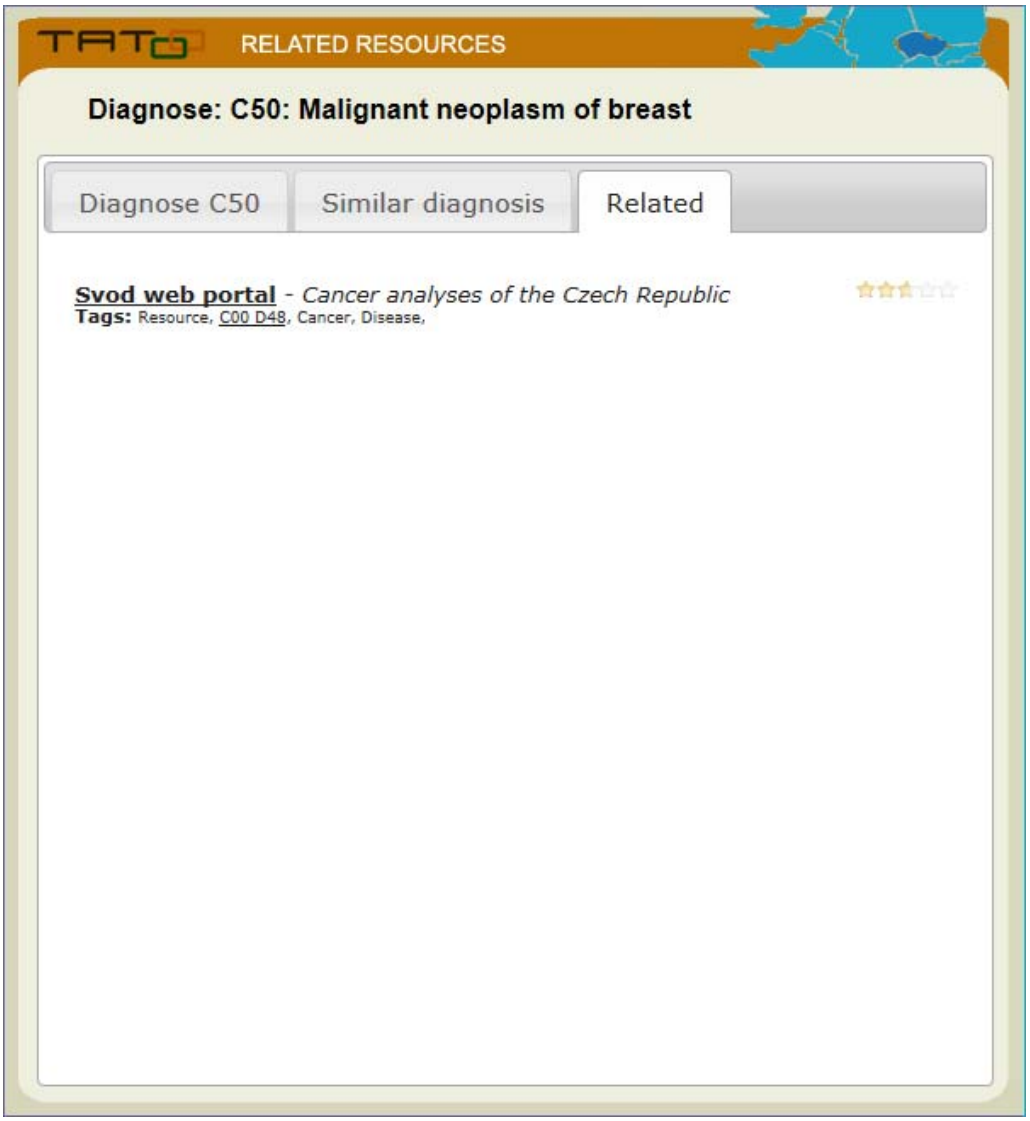

Figure 12. List of related resources in SVOD portal (live demo)

5. Conclusion

The Validation Scenario - Case 3, which studies the anthropogenic impact and influence of global climate change the transport and fate of persistent organic pollutants (POPs) and their impact on human health, aims to improve the discovery process of two domains (POPs, cancer) and to find connections between them. It profits from the GENASIS and SVOD portals (created

at MU), which offer access to the data required to perform such analyses. This document describes how the Validation Scenario profits from these two portals, but additional resources can be easily added and incorporated to enlarge the breadth of the analysis.

The following Use Cases have been described:

- UC1: Discover resources with existing tools
- UC2: Generic discovery
- UC3: Persistent Organic Pollutant Resource Discovery
- UC4: Oncological resource discovery
- UC5: Define resource uncertainty
- UC6: Compare resources
- UC7: Find similar resources
- UC8: Find related resources

Validation is composed by a set of tests based on the above Use Cases that verifies the correct deployment as well as the successful invocation of TaToo's Public Services. Based on the different nature of the TaToo software, the validation varies between portal system tests and services and components tests. The portal system tests will be grouped into the Positive User Testing (perform correct operations on the portal, generating coherent final results) and the Negative User Testing (perform invalid operation and / or provide invalid user inputs).

Further, main objectives of testing are:

- Ensure correct behaviour of the TaToo Semantic Search and Discovery Framework;
- Measure scalability and stability of the system;
- Measure system performance;
- Identify issues occurred during the validation process.

As this document is published (April 2012), Use Cases UC1, UC7 and UC8 are ready as a live demo for testing by experts, these demos are available on non-public parts of SVOD and GENASIS portals. Use Cases UC3 and UC4 has been implemented as a part of TaToo Portal and are available for internal testing.

5.1. Acknowledgements

"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."

6. References

Baeza-Yates, 1999 Baeza-Yates R. et al.: Modern Information Retrieval. Addision Wesley . ISBN 0-201-39829-X

Dusek, 2005 DUŠEK Ladislav, MUŽÍK Jan, KUBÁSEK Miroslav, KOPTÍKOVÁ Jana, ŽALOUDÍK Jan, VYZULA Rostislav. Epidemiology of Malignant Tumours in the Czech Republic (online). Masaryk University, Czech Republic, 2005, [cit. 2010-6-25]. http://www.svod.cz. Version 7.0, 2007, ISSN 1802 – 8861.

Dusek, 2009 Dušek, L., et al. : Czech Cancer Care in Numbers 2008-2009. Grada. ISBN 978-80-247-3244-2

Pinchot, 2000 Pinchot H., The Clinical Stage: Its Definition and Importance in Prostate Cancer, April 2000 v3.1, http://www.prostate-cancer.org/education/staging/Pinchot_Clinical_Stage.html

McDevitt, 2010 McDevitt J., Comber H.: A proposed CORE national Cancer Dataset: National Cancer Registry, January 2010, v1.0, http://www.ncri.ie/news/20100118.shtml

Dem, 2007 Demografická příručka 2007. Český statistický úřad, 2008. Dostupné z http://www.czso.cz/csu/2008edicniplan.nsf/publ/4032-08-2007

NOR, 2007 ÚZIS ČR (2007). Ústav zdravotnických informací a statistiky ČR, Národní zdravotnický informační systém (NZIS), Národní onkologický registr (NOR), 20.12.2007, dostupné z http://www.uzis.cz/info.php?article=368&mnu_id=7300

Chen, 1976 Chen P.P., "The entity-relationship model: toward a unified view of data", ACM Transactions on Database Systems 1:1 pp 9-36, 1976

