SVOD – System for Visualizing of Oncological Data and Their Semantic Enhancement

Miroslav Kubásek, Jiří Hřebíček, Ladislav Dušek, Jan Mužík, Jiří Kalina

SVOD - System for Visualizing of Oncological Data and their semantic enhancement

Miroslav Kubásek, Jiří Hřebíček, Ladislav Dušek, Jan Mužík, and Jiří Kalina

Institute of Biostatistics and Analyses, Masaryk University,
Kamenice 126/3, 625 00 Brno, Czech Republic
{kubasek, hrebicek, dusek, muzik, kalina}@iba.muni.cz

Abstract. We present an overview of the SVOD web portal (System for Visualizing of Oncological Data) which is focused on population risk analyses related to cancer epidemiology and show its integration with the FP7 project TaToo (Tagging Tool based on a Semantic Discovery Framework). The developed TaToo Tools provide a semantic web solution in order to close the discovery gap that prevents full and easy access to specific web resources. Furhter more we will discuss in detail the used ontological resources, the integration of TaToo Tools into the SVOD portal together with the evaluation of this integration.

Keywords: cancer, epidemiology, ontology, SVOD, TaToo, web portal

1 Introduction

Cancer epidemiology can be regarded as one of the most important and most frequently analysed topic in the field of human risk assessment. It is not only due to a remarkable public concern about growing population risk, cancer incidence and mortality is evident and a clearly reachable endpoint for risk assessment studies. We can enter this problem from the viewpoint of risk factors as agents initiating carcinogenesis, but epidemiological parameters can retrospectively indicate hazardous impact on population on a large scale. The bioindication from epidemiological data of course requires sufficient data sources. It means representative long-term profiles of incidence and mortality and a very good awareness of the most important risk factors.

In order to analyse epidemiological trends we must be able to distinguish between statistically significant trends and random fluctuations. For risk assessors it would be most important to recognize the environmentally related cases that can be attributed to external factors like pollution of air, drinking water and/or food. And again, the influence of these factors must be filtered on substantial background of the other "natural" risk factors like the age structure of the population, genetic factors or the frequently omitted life style. Apart from these complicated circumstances we can concentrate our attention to some cohorts of oncological patients that might probably indicate an external impact – if they increase in incidence and non-randomly in some regional or temporal scales. It means namely the occurrence of less advanced disease stages in age groups that are commonly out of the main risk (for example the breast carcinoma

cases in pre-menopausal women younger than 35 – 40 years). In that context, we can take some information benefit mainly from diagnostic groups that might be related to several external influences (colorectal carcinoma, kidney and lung carcinoma, breast cancer, malignant melanoma, …).

All these analyses require easily available large data sets that are themselves very expensive and typically not directly available, e.g. epidemiological cancer registries. In addition to this, we must aggregate cancer data with demographic data in order to obtain for example age-specific profiles of incidence. That is why there is such a growing interest of many professional groups (health care managers, environmental experts, risk assessors) in accessibility of these data. In our experience however, the demand for data cannot be easily fulfilled by blind databases of primary population data. The analyses are very time-consuming and finally, the outputs might be ambiguous and not assured enough to be communicated to the public. Therefore, we developed a professional web portal SVOD (System for Visualizing of Oncological Data) that offers automatically generated and verified epidemiological analyses on cancer incidence and mortality in Czech Republic [1]. This presentation is aimed to introduce potential users to the technology, architecture and functionality of the portal. The SVOD portal is accessible through the following web address: http://www.svod.cz. The portal user interface can be switched between Czech and English language.

2 System for Visualizing Oncological Data

Creating 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 a wide spectrum of users. We anticipate that general epidemiology data about these serious diseases and related population risks should be freely available to everybody in the Czech Republic [2]. Another ambition of this web portal is to provide relevant information about tumour epidemiology in the Czech Republic abroad [3].

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

1. *Current news*: regularly updated information about population risk assessment and tumour epidemiology;
2. *Interactive analyses* that allow the user to investigate 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 secured, validated and publishable data of tumour epidemiology, without any privacy data of patients.

The project of creating the web portal SVOD for tumour epidemiology in the Czech Republic is tied to a longstanding development of analytical software for data from the National Oncological Registry (NOR) [4]. It was created in the years 1999-2003 and allows access to all NOR data via a wide range of automated analyses. Although the information and communication technologies (ICT) of SVOD were finalized successfully, there are severe limits with its distribution and availability. The

web portal SVOD solves all these problems and provides an effective way of accessing epidemiological analyses by an unlimited number of users.

The SVOD portal processes mainly data from NOR [4] which is managed by the Institute of Health Information and Statistics (UZIS CR)[1]. It offers validated epidemiological data from the years 1977 - 2010. This represents a unique data set (of 1.6 million records) at least in the European region.

Epidemiological trends cannot be produced without relevant demographic data about the examined population. This data was provided from the Czech Statistical Office (CSU)[2] on the basis of general cooperation agreement with the Masaryk Memorial Cancer Institute[3] in Brno and the Masaryk University (MU) in Brno.

The SVOD portal was created by the Institute of Biostatistics and Analyses at the Faculty of Medicine and the Faculty of Science of the Masaryk University in Brno and Masaryk Memorial Cancer Institute in Brno [1-2]. The development of the SVOD portal is vitally supported by the Ministry of Health of the Czech Republic in the context of the National healthcare quality programme. Further development was supported by a research programme of the Masaryk Memorial Cancer Institute (Functional diagnostics of tumours, MZO 00209805) and a research programme of Faculty of Science MU (INCHEMBIOL - RECETOX project, No. 0021622412[4]). These grant projects guarantee a long-term viability of the SVOD portal and ensure regular updates of data and a successive development under supervision of administrators.

Information services of the SVOD portal 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 (like the cooperation with above mentioned INCHEMBIOL project). The current version of the SVOD portal offers only epidemiological data, but NOR database allows even analyses of diagnostics and treatment of oncologic patients and survival analyses - even in relation to the current hospital. All these analyses are prepared only for communication in the Oncological Society[5] and are available in a restricted zone of the portal. It therefore serves as an information source for Czech health management and helps to set up reference standards for healthcare results in oncology [2].

3 Web-based analytical tools

Interactive analyses are one of the core functions of the SVOD portal. Through these functions, you can easily analyze epidemiological trends examined over three decades, stratify and filter cohorts of patients and extract population risk in absolute or age-specific values. Some of these analyses offer benchmarking with respect to the

[1] http://www.uzis.cz
[2] http://www.czso.cz
[3] http://www.mou.cz/en/
[4] http://www.recetox.muni.cz/inchembiol/index.php?language=en&id=
[5] http://www.linkos.cz/

clinical status of the disease. Major epidemiological trends are available in comparison with international data - GLOBOCAN 2008 [5].

3.1 Incidence and mortality analyses

These analyses calculate time trends of crude incidence, crude mortality and mortality/incidence ratio. Available parameters for calculations are absolute numbers, crude rates (number of cases per 100000 people in population) and age standardized ratios (ASR - European or World standard). User can also specify some filters (constraints) for analyse e.g. sex and age of patients, specify region, time period, stage, TNM and other parameters through the buttons available in the 'Incidence and Mortality' panel (see **Chyba! Nenalezen zdroj odkazů.** - left). These filters are available for all analyses, but some filters are disabled depending on the concrete analyses.

3.2 Regional overview analyses

These analyses return a comparison of incidence and mortality in regions of the Czech Republic. Available parameters are the crude rate and the age standardized ratio. The output could be obtained as bar-plots or maps. The area filter for this analyses is disabled (see **Chyba! Nenalezen zdroj odkazů.** - right).

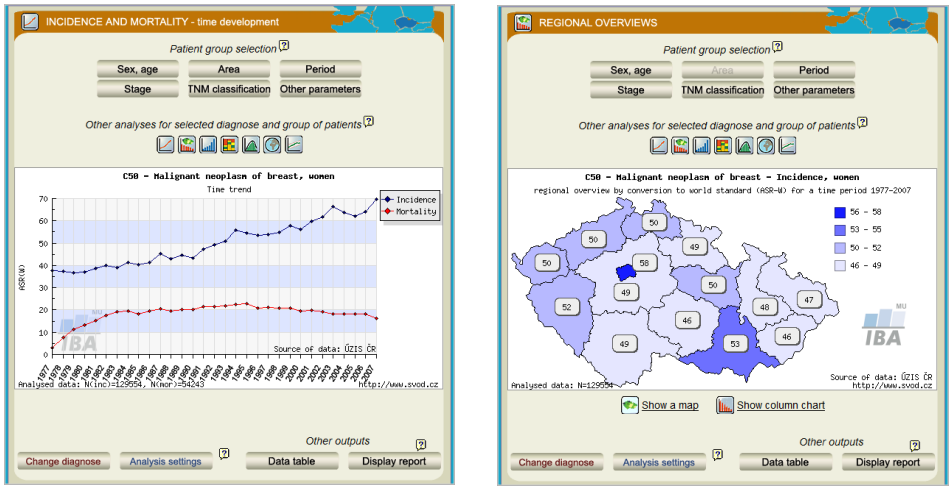

Fig. 1. Example of incidence and mortality and a regional overview analyses. (Source: Authors)

3.3 Time trends

Time trends describe changes of incidence and mortality with respect to time. Available parameters are growth indexes related to a selected year and between-years changes. Both parameters could be viewed as absolute numbers or as relative proportions. The user can select the reference year for the analyses.

3.4 Clinical Stage analyses

The Clinical Stage analyses represent time trends of the proportion of patients in specific clinical stages. Available parameters are absolute numbers, % and crude rate of patients in specific clinical stages. Available outputs are time trend bar-plots, time trend line-plots or pie charts of the selected time period. The Stage and TNM classification filters are disabled for this analyses.

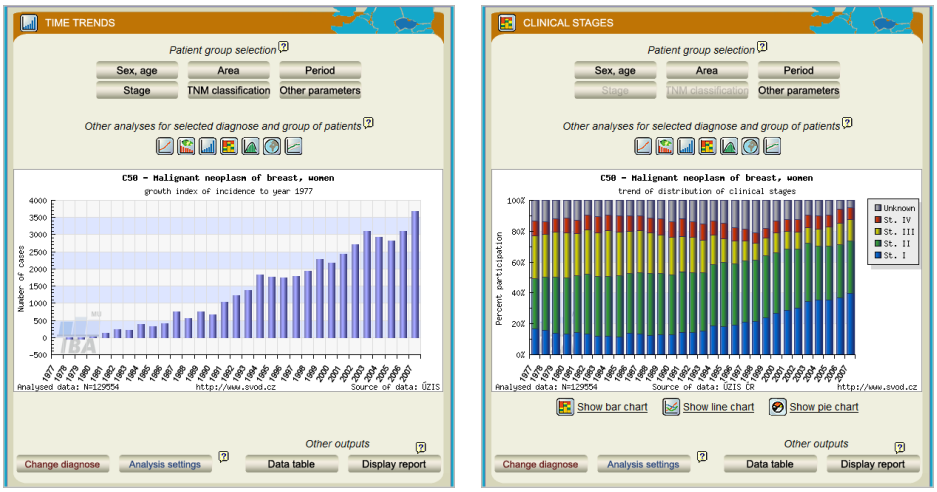

Fig. 2. Example of time trends and clinical stages analyse (Source: authors)

3.5 Age analyses

These analyses return the age structure of population of patients with a selected diagnose. Available parameters are the absolute numbers of cases in age categories, % of cases in age categories and the age specific rate (number of cases in age category per 100000 people in population cohort of the same age).

3.6 International data analyses

These analyses return a comparison of incidence and mortality in the Czech Republic (CZ) with other countries. All these analyses are based on data obtained from the IARC database GLOBOCAN 2008 [5]. Comparative standards: time trend of incidence or mortality in a selected region in comparison with the situation in the whole Czech Republic. Available parameters are the crude rate and the age standardized ratio (ASR - European or World standard).

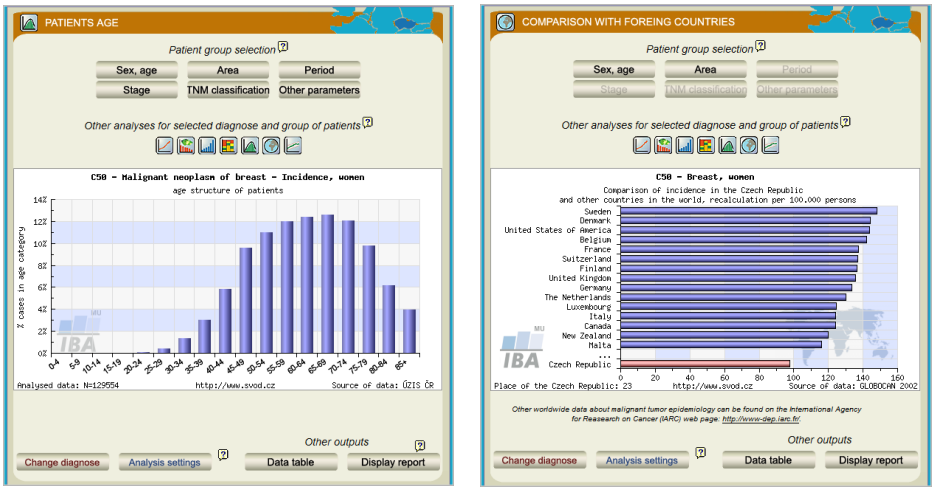

Fig. 3. Example of age analyses and in comparison with foreign countries (Source: authors)

4 Integration with TaToo Tools

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 [6]. 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.

Integration of TaToo Tools into SVOD portal is part of the Masaryk university[6] validation scenario (MU scenario), which is the anthropogenic impact and the influence of global climate change on this impact [7-9]. The purpose of this chapter is to give the overview of implementation of this integration.

4.1 Inventory of Ontologies

First, we briefly describe the knowledge resources related to the cancer and epidemiology domain. These resources vary from ontological resources to non ontological including descriptions of technologies, existing ontologies or concrete vocabularies.

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

By analysing the cancer and epidemiology domain we finally proposed the initial version of ontology. For modelling of the ontology we used the Web Ontology Language (OWL) and the NeOn Toolkit[7]. Our proposed ontology contains 351 concepts and was used as the base concept for describing (tagging) resources from the cancer and epidemiology domain .

ICD-10 Ontology

The International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) [11] is a coding of diseases and signs, symptoms, abnormal findings, complaints, social circumstances and external causes of injury or diseases, as classified by the World Health Organization (WHO)[7]. The code set allows more than 14,400 different codes and permits the tracking of many new diagnoses. Using optional sub-classifications, the codes can be expanded to over 16,000 codes. Using codes that are meant to be reported in a separate data field, the level of detail that is reported by ICD can be further increased, using a simplified multiaxial approach. The International version of ICD should not be confused with national Clinical Modifications of ICD that include frequently much more detail, and sometimes have separate sections for procedures. Work on ICD-10 began in 1983 and was completed in 1992.

Disease ontology

The mission of the Disease Ontology (DO) is to provide by open source ontology for the integration of biomedical data that is associated with human disease. DO will have a formally correct (in the ontology sense), semantically computable structure. Terms in DO will be well defined, using standard references. These terms will be linked to well-established, well-adopted terminologies that contain disease and disease-related concepts such as SNOMED[8] (we are working with SNOMED to see if we can release SNOMED codes linked via UMLS[9] to the community), ICD-10, MeSH[10], and UMLS. The combination of a semantically computable structure and the external references to these terminologies will enable useful inference between disparate datasets using one or more of these standard terminologies to code disease. The Disease Ontology will be, at the end of this project, a community-driven, community-accepted ontology of diseases for clinical research and medicine inclusive of genetic, environmental and infectious diseases. It will encapsulate, therefore, a comprehensive theory of disease. The design of the disease ontology will enable greater understanding of disease states by placing heritable disorders in the context of other infectious diseases and related diseases. The structure of Disease Ontology and the external references to other terminologies will enable the integration of disparate datasets through the concept of disease.

---

7 http://www.who.int/en/

8 http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

9 http://www.nlm.nih.gov/research/umls/

10 http://en.wikipedia.org/wiki/MeSH

TNM ontology

We did not find an overall ontology for the structuring of TNM classifications, but there are few research initiatives in this field. One ontology which can be taken into account for creating an ontology is "An Ontology for Carcinoma Classification for Clinical Bioinformatics " by Anand Kumar et. al.[12] . They outline a formal theory for addressing these issues in a way that inferences drawn upon the ontologies would be helpful in interpreting and inferring on the entities which exist at different anatomical levels of granularity. Their case study is on the colon carcinoma, one of the commonest carcinomas prevalent within the European population.

4.2 Ontology Enhancement of Analytic Tools

The purpose of ontology enhancement of SVOD portal analytic tools is to provide the users of the SVOD portal with the possibility to indirectly use the TaToo functionality for the discovery of similar resources based on analysed objects. The TaToo discovery can be started directly from the web analyses tools. The relevant information needed for the search is already entered via the web interface during the analyses and can be submitted to the TaToo Framework.

In this integration we implemented a "TaToo button" (see Figure 5) into SVOD portal analytic tools and applied the TaToo Discovery Service to retrieve a list of related data resources from the TaToo Repository (see Figure 4). Queries used in calling the Discovery Service are automatically built from the actual settings of the SVOD analyses tool. Communication between TaToo Discovery Service and developed discovery application is based on SOAP.

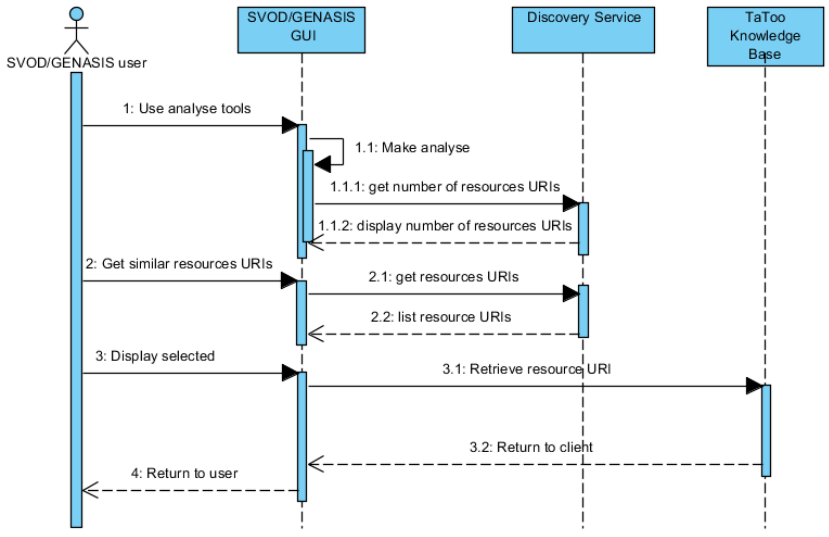

Fig. 4. Sequence diagram of communication between SVOD portal and TaToo services (Source [13])

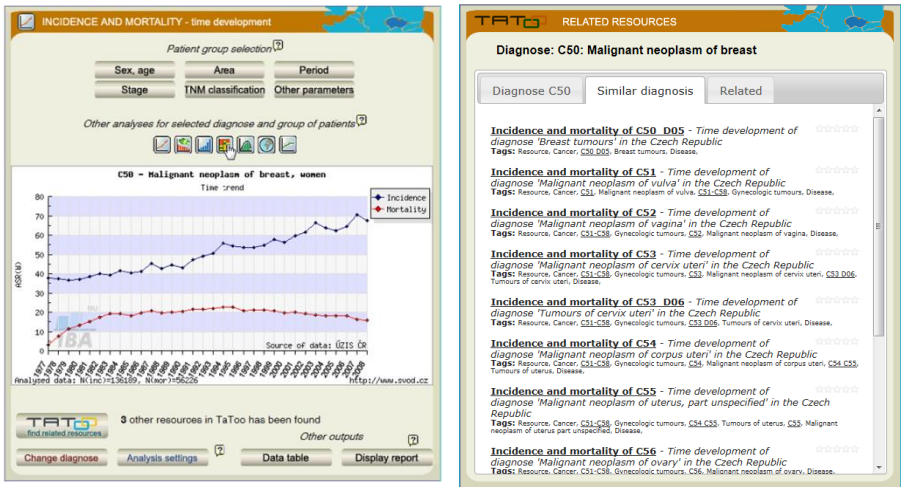

Fig. 5. GUI of integration of TaToo functionality into SVOD portal analytic tool

4.3 Evaluation

The objective of the evaluation of TaToo Tools was to ask external users whether they agree that the objectives of Enhancement of analytic tools have been achieved and how satisfied they are with the provided functionalities.

The evaluation process was organized on three levels depending on the availability of the of the evaluators and additional three workshops. The first group “A” (9 persons) had been invited to three workshops. Two of them were organised during 8th and 9th summer schools of applied informatics in September 2, 2011 and September 7, 2012, where all participants were introduced to the Validation Scenario “Anthropogenic impact and global climate change” and the basic idea of TaToo Project. The discussion and feedback of participants were positive and they obtained public access to the TaToo and SVOD portal. After this introduction in two workshops, we gave all participants of the third workshop, held on November 26, 2012, credentials to access both the TaToo portal and the SVOD portal application which integrated the TaToo Services. We provided them a questionnaire for collecting their feedbacks. This group was given one week's time to get everything tested and sent back thereafter their questionnaires.

The questionnaire has been structured as follows: Two statements regarding relevance and usability could be rated on a five-point scale ranging between 1 ("I agree") and 5 ("I disagree"). The statement regarding relevance is "The function ... is interesting for my work", the statement regarding usability is "The function ... is user-friendly". Furthermore, there was an open question asking for further comments and proposals how the application could be improved.

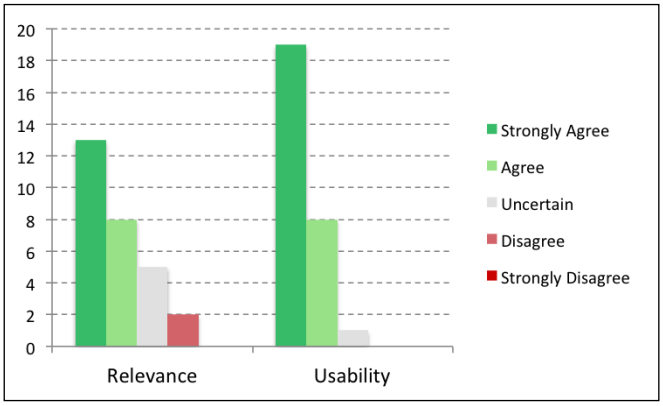

Fig. 6. Assessment of extension of SVOD portal (Source [13])

Overall, the idea of tagging resources through the topics of interest was evaluated very positively by the domain users. Linking TaToo with the other specialized portals was recognised as a good idea, but the problem may be in the fulfillment of the repository. TaToo portal users perceive as a working tool, but for its practical application are sceptical. To administer the repository it would probably be better to use some easier and user-friendly solutions. It would also be appropriate for the end user to hide the ontological background of the TaToo Tools and offer only understandable (e.g. more human readable) concepts.

5 Conclusion

We have introduced the SVOD (System for Visualizing of Oncological Data) portal and described the vision of its integration and ontological enhancement by the FP7 project TaToo. The aim of TaToo is providing a collaborative platform for the semantic enrichment of environmental information resources on the web. The development of domain ontologies in the TaToo Ontology framework is an evolutionary process which will be extended by adding new specific concepts and relationships that are currently not provided. We also gave a short overview of realized functionalities, the software quality evaluation and the usability evaluation of the user interface. A number of domain experts were asked to assess the relevance and usability of SVOD tools and its TaToo extension. We can conclude that the collected feedback was predominantly positive.

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. 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)
2. Dušek, L., Hřebíček, J., Kubásek, M., Jarkovský, J., Kalina, J., Baroš, R., Bednářová, Z., Klánová, J., Holoubek, I.: Conceptual model enhancing accessibility of data from cancer–related environmental risk assessment studies. In: Hřebíček, J., Schimak G., Denzer, R. (eds). 9th IFIP WG 5.11 International Symposium on Environmental Software Systems: Frameworks of eEnvironment, ISESS 2011. pp. 461-479 Springer, Heidelberg (2011)
3. Dušek, L., at al. Czech Cancer Care in Numbers 2008–2009. Grada Publishing, Praha (2010)
4. Czech National Cancer Registry, http://www.uzis.cz/ang/ccr/ccrindx.htm (2013)
5. GLOBOCAN 2008: Estimated cancer Incidence, Mortality, Prevalence and Disability-adjusted life years (DALYs) Worldwide in 2008, http://globocan.iarc.fr/ (2008)
6. 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)
7. 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 Hřebíček J., Pitner T., Ministr J. (eds). 7. Summer school of applied informatics. Indikátory účinnosti EMS podle odvětví. pp. 6-23, nakladatelství Littera, Brno (2010)
8. 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)
9. Kubásek, M., Hřebíček, J., Kalina, J., Dušek, L., Urbánek, J., Holoubek, I.: Global Environmental Assessment Requires Global Functional Searching Engines: Robust Application of TaToo Tools. In Hřebíček, J., Schimak G., Denzer, R. (eds). 9th IFIP WG 5.11 International Symposium on Environmental Software Systems: Frameworks of eEnvironment, ISESS 2011. pp. 507-518. Springer, Heidelberg (2011)
10. Gómez-Pérez A., Motta E., Suárez-Figueroa M.C., NeOn Methodology in a Nutshell, http://www.neon-project.org/nw/NeOn_Book (2010)
11. World Health Organisation. ICD-10 Classifications of Mental and Behavioural Disorder: Clinical Descriptions and Diagnostic Guidelines. Geneva. World Health Organisation. http://www.who.int/classifications/icd/en/ (1992)
12. Kumar A, Yip YL, Smith B, Marwede D, Novotny D.: An Ontology for Carcinoma Classification for Clinical Bioinformatics, MIE 2005;116:635-40 (2005)
13. Kubásek M., Hřebíček J., et. al.: Deliverable D5.3.9 - Final Implementation and Validation Report - Case 3 – V3, TaToo Project, Public Document, http://www.tatoo-fp7.eu, 2012