International Environmental Modelling and Software Society (iEMSs)
2010 International Congress on Environmental Modelling and Software
Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada
David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.)
http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings

TaToo: tagging environmental resources on the web by semantic annotations

Andrea E. Rizzoli[a], Gerald Schimak[b], Marcello Donatelli[c], Jiri Hrebicek[d], Giuseppe Avellino[e], Jose Lorenzo Mon[f], Ioannis Athanasiadis[1]

[a] IDSIA, Lugano, Switzerland, (andrea, ioannis @ idsia.ch)
[b] AIT Austrian Institute of Technology, Austria (gerald.schimak@ait.ac.at)
[b] JRC, Ispra, Italy (marcello.donatelli@jrc.it)
[d] Masaryk University, Brno, Czech Republic
[e] Elsag Datamat, Italy
[f] Atos Origin, Spain

Abstract: The web is rapidly evolving and its traditional role of repository of static information is changing into a hub for collaboration among people. Web resources tend to become more and more complex, and to offer services that include access to remote databases, and computational power. All of this becomes very interesting not only for the common user, but especially for scientists and researchers which actually see their computers "disappear" into the web "cloud", getting back an unprecedented access to services and computational resources.

Yet, to exploit these new facilities new tools are needed. The TaToo project aims at exploiting a common practice among web user: search, discovery and tagging of interesting resources. The practice of tagging allows user groups to label and classify resources enabling aggregators to display the most relevant ones according to the context. TaToo aims to take the core idea of tagging and adding the ability to add valuable information in the form of semantic annotations, thus facilitating future usage and discovery, and kicking off a beneficial cycle of information enrichment. Thus, the production of semantic meta information will improve the discovery process, but also interpretation in a larger sense (verification that its the information I was looking for, assessment of usefulness for a given situation, understanding of how to use the information correctly etc.).

***Keywords**:* semantic annotation; semantic tagging; model search and discovery; web services; environmental information enrichment.

1. Introduction

Technological progress has been constantly providing new tools and techniques for environmental scientists. If we look back at the not so distant past, geologists, geographers, hydrologists, ecologists and other environmental scientists they all had to manually and painstakingly collect data by surveys, which were expensive, labour intensive, and required lots of time. Data were then stored in paper based repositories and archives, and processing data was also limited by the availability of data. Perhaps the most brilliant use of spatial data and analytic reasoning was the discovery of the source of a cholera epidemic in

London by Dr John Snow. In 1854 he identified the water pump of Broad Street in London by plotting deaths on a map of the borough, as represented in Figure 1.

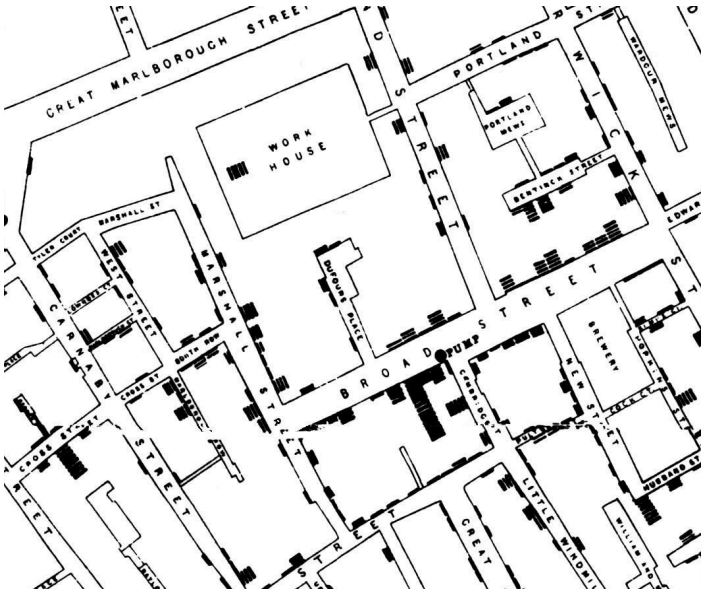

Figure 1. A section of the original map published by Dr Snow in 1854. The black bars represent deaths at a given house number.

This type of representation would require only a few minutes work using a modern GIS tool, but it took Dr Snow several days of work to put together the above map, and a considerable effort in collecting the data.

Nowadays data can be automatically collected by remote sensing or by sensor networks; they are stored in information systems relying on advanced database techniques and data storage facilities. Data representation is facilitated by desktop Geographical Information Systems (e.g. ArcGis, MapWindows) or even web-based GIS-like applications (Google Maps, Bing Maps). Data processing is provided by sophisticated and complex models, supported by advanced computer architectures, exploiting distributed and parallel processing.

Given the state of things, the outlook for the environmental scientist should be particularly bright and rosy: vast amounts of data to process, a great variety of models available to process and elaborate data, and powerful visualisation tools. While we must regard with awe what we have achieved in the past 30/40 years, we should also be wary of the threat posed by having access to too much information, which we cannot neither discern nor make sense of.

Scientists and researchers have been aware of such a threat for a long time, and various approaches and mitigation measures have been devised. We can enumerate a few: data catalogues, metadata, model bases and model repositories (such as EIONET, UN FAO, etc.). At the same time, the pressure towards the integration and exchange of data pushed towards the creation of standards for data representation, such as HDF (hierarchical data format), the OGC standards for GIS data, and standards for model integration, such as OpenMI as described in Gregersen [2007].

Standards and metedata are part of the cure, but they are still too little to front the exponential increase in data availability provided by earth observation initiatives such as INSPIRE[1], Google Earth[2], OpenTopography[3], GEOSS[4] and many others.

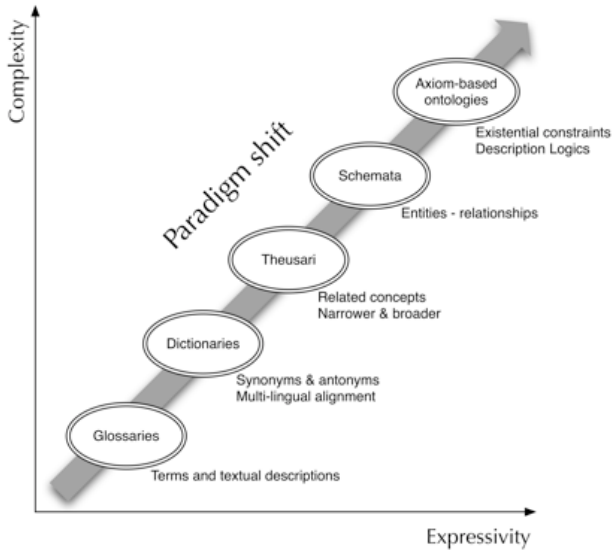

Figure 2. From glossaries to axiom-based ontologies: a wide palette of choices to annotate resources at increasing levels of complexity and expressivity (from Villa et al. 2009).

We claim that the expressivity of glossaries, dictionaries, thesauri and schemata is too limited for the demands that we expect to be posed to environmental information systems in the near future.

The Internet is growing in a non-coordinated manner, where different groups continuously publish and update information, adopting a variety of standards, according to the specific domain of interest: from agriculture to ecology, from groundwater to climate change. This unconstrained and unregulated growth has proven to be very successful, as more information is made available, even more is being added, in a virtuous cycle of information accrual. At the same time, modern search engines make looking for information rather easy, and despite some studies reprieve the presence of non-relevant hits in the result of most queries (e.g. Gordon and Pathak [1999]), we personally judge the overall performance more than satisfactory for most users (we don't have data to support our claim, but if search engines were so bad, probably people would not use them...).

Yet, searching and discovering information requires a good deal of expertise and pre-existing knowledge. That may not be a problem when the user is searching its own domain of expertise, but what happens when the user is trying to gather environmental information on a trans-disciplinary study? Or when preparing and integrated assessment study? These type of research and development efforts require the detailed knowledge of multiple domains, from economy to ecology, from hydrology to social sciences. In our experience, we have seen situations where different groups of scientists labelled the same concepts with different terms, and also labelled similar (but distinct) concepts with the same names (Athanasiadis et al. [2009], Janssen et al. [2009]).

---

[1] http://inspire.jrc.ec.europa.eu

[2] http://earth.google.com

[3] http://www.opentopography.org

[4] http://www.earthobservations.org/geoss.shtml

To overcome this type of problems we need to add rich semantics, possibly axiom-based semantics, to our environmental resources, thus increasing the expressivity of information, but, at the same time, also increasing the complexity required to convey it, as shown in Figure 2 in Villa et al. [2009].

The recently started (beginning of 2010) TaToo project tries to fill this information and discovery gap, by providing a way to semantically annotate environmental resources on the web. The project idea is strongly inspired by existing social bookmarking initiatives, such as Delicious, reddit, StumbleUpon, Digg etc. Yet, TaToo aims to let user use semantics in their annotations, by accessing shared ontologies, thus enabling inference engines to process information and discover new facts and new relationships that are not explicitly stated in the body of knowledge.

In the remainder of this paper we will first review the state of the art in semantic annotations and tagging, then we will present our vision of how TaToo should work and operate, and we also describe the preliminary draft of the enabling software architecture. Finally we introduce some test cases, which will be developed during the project, and finally we will conclude by outlining our expectations for the future of this project, that is expected to end in three years time.

2. The semantic web and semantic annotations

The Semantic Web vision aims to provide mechanisms to augment the Web content with semantic annotations and retrieve concrete information contained in the Web, not only references to pages or resources where the requested information could be contained within. Semantic Web technologies have matured considerably in the last years and standardisation efforts have created standards for data exchange (Resource Description Format (RDF[5])), ontology languages (e.g. Web Ontology language (OWL[6])) and it is progressing in the standardisation of further components of the Semantic Web such as SAWSDL[7], a language allowing Web services to be described with semantics; to be more amenable to automated processing on behalf of the users. Another relevant component is GRDDL/RDFa[8]: a way to embed RDF-based annotation into existing Web pages.

While the progress in standards for the Semantic Web is highly relevant to support the annotation of unstructured information contained in web pages, we are particularly interested in the annotation of web services (see the work of Hilbring and Usländer, [2006]), as most environmental resources will be made available under that form. Everything, from a model run, to a database query, or a GIS operation, can be served as a web service, and there are impressive efforts towards the standardisation of web services in the environmental domain, such as OGC's OpenGIS web service common standard[9].

For these purposes we refer to the WSMO[10] conceptual framework for the description of Web Services currently realised in the WSML family of languages and the reference implementation WSMX. The WSMO Lite language realises the WSMO concepts using

---

[5] http://www.w3.org/RDF/

[6] http://www.w3.org/TR/owl-features/

[7] http://www.w3.org/2002/ws/sawsdl/

[8] http://www.w3.org/2004/01/rdxh/spec

[9] http://www.opengeospatial.org/standards/common

[10] http://www.wsmo.org

RDF/S and SAWSDL. WSMO Lite can thus be used as the basis for a standard W3C ontology for modelling Web services.

Of interest are also DAML-S, OWL-S that standardise semantic web services OWL-S (formerly DAML-S) is a semantic mark-up language for Web services. It is an academic ontology proposal, submitted to the W3C, for describing Web services based on OWL. It provides a core set of constructs (machine-understandable) to provide descriptions of Web services properties and capabilities. Through OWL-S support, users or agents acting on behalf of the users (including software agents) are able to discover, invoke, compose, and monitor Web resources. Even if OWL-S has not reached yet the status of a standard, it is quite stable and adopted. There is still a good amount of research going on that involves OWL-S mainly because it is able to address workflow management issues.

While a thorough state of the art analysis cannot be included here for sake of the limited available space, we can simply state that a great number of projects and initiatives are working towards the semantic annotation of web services, such as SWWS[11], DIP[12], SOA4ALL[13], while others are focusing more on the discovery and retrieval of annotated content, such as INFRAWEBS[14], OWL-S Matchmaker (Sykara et al, [2003]), BREIN[15] and again SO4ALL. The aim of TaToo is to join these lines of work and to contribute with the development of environmental-specific applications of the above-mentioned research topics and technologies.

3. The vision

Despite the great amount of work and resources currently deployed in the field of semantic annotation of web resources, there are some major hurdles to be overcome to make the TaToo vision become a reality.

TaToo is expected to work along the lines of one of those social tagging and bookmarking website. Here we focus on two major use-cases: finding and annotating a resource, and searching for and discovering an annotated resource.
In the first case, the user stumbles on an interesting resource during his/her work. Here we assume that the resource is a web service described in a web page. The user has simply to feed the URL of the web page to the TaToo server application. This can be done by simply dragging and dropping the URL in a modern browser such as Firefox. The TaToo server application recognises the URL and starts processing the page, automatically extracting the information regarding the web service and processing the text in the webpage. The TaToo server application then generates a web page where the various elements of the original webpage are presented to the user and offered for semantic tagging. The process is therefore a mix of automated semantic annotation, and manual annotation. The user-friendliness of the interface will be therefore a critical element for the success of the application.

In the second case, the user accesses the TaToo server to search for and discover environmental resources which have been semantically annotated. The advantage is the ability to come across the limitations imposed by specific domain jargons and semantic

[11] http://swws.semanticweb.org

[12] http://dip.semanticweb.org

[13] http://www.soa4all.eu

[14] http://www.infraWebs.eu

[15] http://www.eu-brein.com

ambiguity. In section 5 we describe some case studies where this feature will be exploited. Also, in a possible future scenario, third-party web-services will be able to use TaToo discovery services to automatically chain web-services to answer complex and structured queries, requiring the integrated runs of multiple environmental resources.

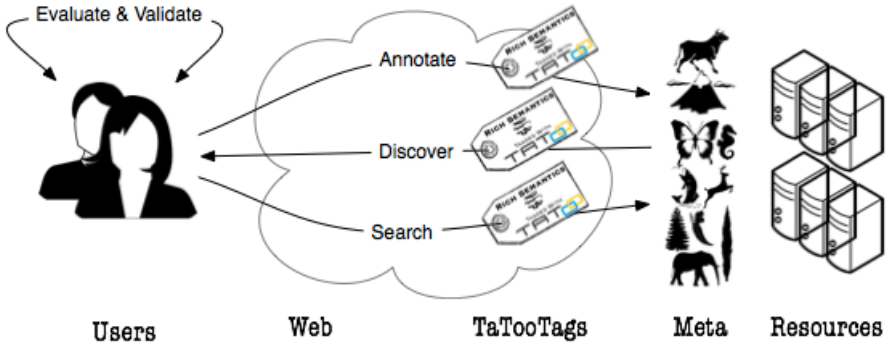

Figure 3. Use cases for the TaToo vision.

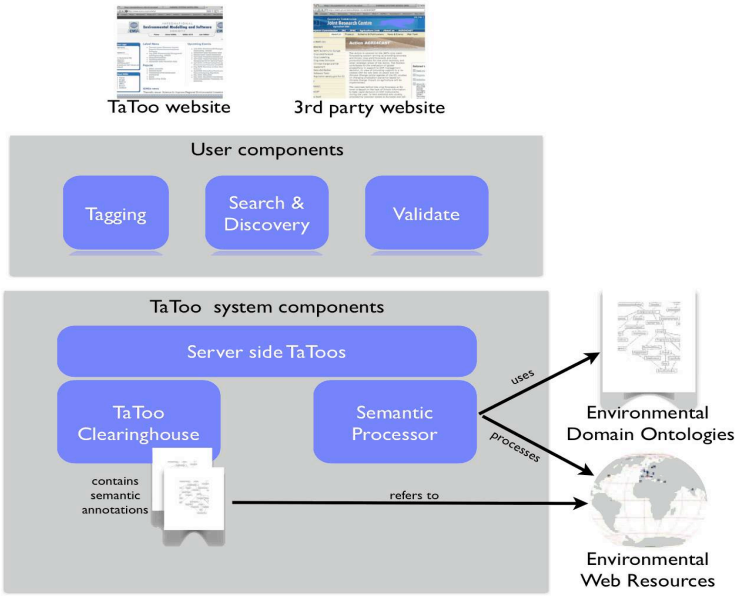

Figure 4. The TaToo overall architecture.

4. The proposed software architecture

TaToo aims to deliver a web based system centred on a three main elements:

- a *clearinghouse*, which plays the role of organising the semantic information on environmental resources. The clearinghouse contains a list of semantic annotations of environmental resources, referring to the original content as available on the World Wide Web.
- a *semantic processor*, a core component of the TaToo system, since it uses a set of (pluggable) environmental ontologies to provide semantic services to the tagging tools of TaToo.

- a set of *tagging tools* (TaToos), which offer services such as tagging new environmental resources, including quality and uncertainty information, searching for and discovering tagged environmental resources, validating the results of a search.

The server side interacts with web based client components. These components can be interactively combined in order to enable human users to create, modify, delete and update TaToos. The composition is open, so that third party will be able to exploit the components to deliver their specifically targeted services. A schematic representation of the architecture is represented in Figure 4.

5. Expected results: the case studies

TaToo plans to validate the usability of its approach through the implementation of three different 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. The scenarios are tackling the following environmental domains: climate change, agriculture, and anthropogenic impacts of pollution.

In the *climate change* case study the aim is to be able to identify model regions, where the current climate matches with the expected future climate of the source region of interest. We call such region pairs with similar climate conditions (at different times) "Climate Twins". A web-based "climate twins" exploration tool will identify those Climate Twins, where source grid-cell's values representing future climate show high similarity with the current climate grid. To find climatic coincidence seems to be a simple exercise, but the accuracy and applicability of the similarity identification depends very much on the selection of climate indicators and uncertainty ranges. The TaToo platform can provide with tools to facilitate an improved, user-focused climate change resource search, through which end-users will be able to add tags and comment existing resources, reuse tags of other users, and eventually discover and retrieve climate twin-region data, through semantic rich, spatially explicit, user-tailored querying.

In the *agro-environmental* case study we will work in collaboration with the AGRI4CAST action of the Joint Research Centre focuses on the European Commission Crop Yield Forecasting System aiming at providing accurate and timely crop yield forecasts and crop production biomass. AGRI4CAST gets increasing requests for analyses to be run against the weather and soil database which require either new or modified modelling capabilities with respect to the set of models available in the operational system. To achieve this, software implementations of Crop Forecasting System model components target the objective of easy composition, extension and re-use. Though detailed model and software documentation is available, along with scientific papers and reports describing the application of the models, still the discovery of appropriate models to-be-employed for on-demand studies is a monotonous task that requires significant human expert efforts. TaToo will be put to the test as a tool to support the proper annotation of resources by defining attributes, such as description, maximum, minimum and default values, units, and URL. Then its search and discovery capabilities will be put to the test to find alternative modelling solutions, given that each component can make available alternate options for estimating/generating variables.

The *anthropogenic impact of pollution* case study will enable 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 knowledge available. Proper use of these

data requires contextual enhancements, which TaToo will deliver through tagging and enhanced meta-data information description embedded into the appropriate semantic environment.

6. Conclusions

In this paper we have introduced the aims and the vision of the TaToo project, a recently started EU-funded project, which aims at providing a collaborative platform for the semantic enrichment of environmental resources on the web. The main challenges of the TaToo project are the provision of an appealing user interface for the semantic annotation of environmental resources, more specifically web services, and the development of a set of tools to provide a preliminary semantic analysis of the content of web resources, with the ability to access different published ontologies that describe the available knowledge basis. Finally, the most critical factor will be the ability to involve the scientific and research community, which are expected to be the prime users of the platform. This is the main reason that drove us to prepare this work, in order to get an early involvement of the user community and to raise awareness about the scope and aims of our future work.

7. Acknowledgement

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

Gordon, M., and P. Pathak, Finding information on the World Wide Web: the retrieval effectiveness of search engines, *Information Processing and Management*, 35(2), 141-180, 1999.

Gregersen, J.B., P.J.A. Gijsbers, and S.J.P. Westen, OpenMI: Open Modelling Interface, *Journal of Hydroinformatics*, *9*(3), 175-191, 2007.

Hilbring, D., and T. Usländer, Catalogue Services Enabling Syntactical and Semantic Interoperability in Environmental Risk Management Architectures., paper presented at EnviroInfo 2006, Graz, Austria, September 6-8, 2006.

Janssen, S. et al., Defining assessment projects and scenarios for policy support: use of ontology in Integrated Assessment and Modelling, *Environmental Modelling and Software, 24*(12), 1491-1500, 2009.

Sycara K., M. Paolucci, A. Ankolekar, and N. Srinivasan, Automated Discovery, Interaction and Composition of Semantic Web Services, *Journal of Web Semantics, 1*(1), 27–46, 2003.

Villa F., I.N. Athanasiadis, and A.E. Rizzoli, Modelling with knowledge: A review of emerging semantic approaches to environmental modelling, *Environmental Modelling and Software, 24*(5), 577-587, 2009.