Information enrichment using TaToo's semantic framework

Gerald Schimak[1], Andrea E. Rizzoli[2], Giuseppe Avellino[3], Tomas Pariente Lobo[4], José Maria Fuentes[4], Ioannis N. Athanasiadis[2]

[1] AIT Austrian Institute of Technology, Seibersdorf, Austria, gerald.schimak@ait.ac.at
[2] IDSIA, Lugano, Switzerland, andrea@idsia.ch, ioannis@idsia.ch
[3] ElsagDatamat, Rome, Italy, giuseppe.avellino@elsagdatamat.com
[4] ATOS Origin, Madird, Spain, tomas.parientelobo@atosresearch.eu, jose.fuentesl@atosresearch.eu

Abstract. 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, with their overall performance being more than satisfactory for most users. Yet, searching and discovering information requires a good deal of expertise and pre-existing knowledge. That may not be a problem when a user searches for common assets using a generic-purpose search engine. But what happens when the user is trying to gather scientific information across boundaries (e.g. cross different disciplines, cross environmental domains, etc)? This asks for new approaches, methods and tools to close the discovery gap of information resources satisfying your specific request. This is exactly the challenge the TaToo project is heading to.

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

1 Introduction

The TaToo project aims at exploiting a common practice among web user: search, discovery and tagging of interesting resources, focusing on environmental ones. Tagging practices allow communities and user groups to label and classify resources, enriching information and enabling aggregators to display the most relevant ones according to the context, as reflected in Fig. 1. TaToo aims to capitalize on the principles of tagging and by investigating the ability to add valuable information in the form of semantic annotations, facilitating future usage and discovery, and kicking off a beneficial cycle of information enrichment.

Thus, the production of semantic metainformation will improve the discovery process, but also its interpretation in a larger sense (verification that it's the information you were looking for, assessment of usefulness for a given situation, understanding of how to use the information correctly etc.).

Standards and metadata 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 (http://inspire.jrc.ec.europa.eu), Google Earth (http://earth.google.com), OpenTopography (http://www.opentopography.org), GEOSS (http://www.Earthobservations.org/geoss.shtml) and many others.

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 [2]. To overcome this type of problems we need to add rich semantics, possibly axiom-based, to our environmental resources, thus increasing the expressivity of information, but, at the same time, also increasing the complexity required to convey it as addressed in [3].

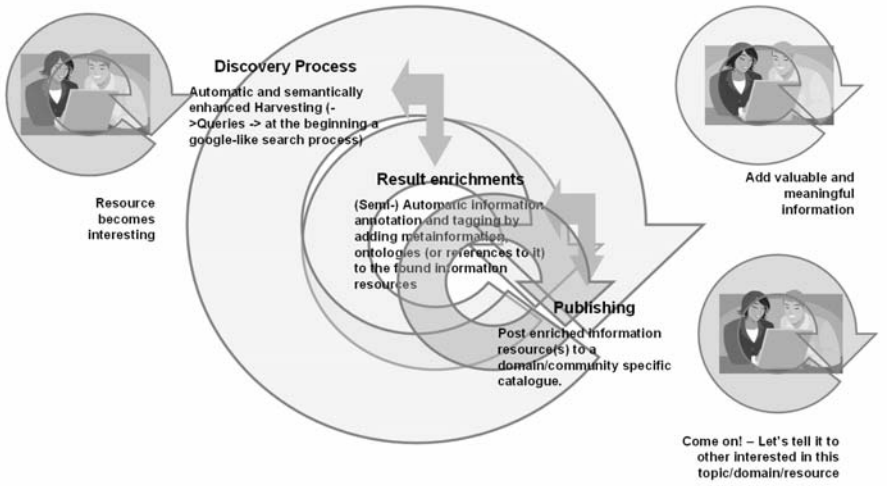

Fig. 1. TaToo's cycles of information enrichment

The TaToo project, which started in 2010, tries to solve this information and discovery gap problem, 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.

TaToo aims for a community driven approach of information enrichments (see Fig. 1), addressing experts, from researchers up to decision makers in authorities within a specific (environmental) communities as well as much as possible the public in order to setup, extent, use and promote their knowledge by using the TaToo framework as a knowledge sharing platform.

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.

2 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. Let's assume (as one of many possibilities) 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 simply by dragging and dropping the URL in a sidebar of a modern browser. The TaToo server application recognizes 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 usability 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 ambiguity. In section 5 we describe some use cases coming form TaToo's validation scenarios 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.

3 The semantic web and semantic annotations

TaToo aims at providing a framework for semantic annotation and discovery of environmental resources. But why use semantic and not informal annotations? The key aspect is precisely that informal annotations are relatively good for human interpretation, but fail to help machines to understand the meaning and perform more advanced tasks based on those annotations. Annotations should be then formal and shared in the domain. In this sense, the Semantic Web [4] was first introduced by Berners-Lee with the aim of providing meaningful web content for machines by bringing formal structure to web resources. Gruber [5] introduced the concept of

ontology as “a formal explicit specification of a shared conceptualization”. Thus ontologies play the role in the Semantic Web as the formal (machine-understandable) and shared (in a domain) backbone. Ontologies are becoming a clear way to deal with integration and semantic discovery of web resources in the environmental domain.

TaToo has the objective of allowing a cross-domain discovery of resources in the environmental field, meaning that resources annotated with different purposes and possibly different ontologies should be retrieved using a common framework. But, there are several challenges to solve this semantic heterogeneity issue:

1. Allowing multi-domain annotation schema: TaToo should allow the annotation of resources using a common and controlled vocabulary based on ontologies. At the same time, TaToo should offer the possibility of particularize the annotations for a given domain.
2. Implementing an extensible discovery mechanism: TaToo should provide a generic discovery framework in order to allow a common way of searching for annotated resources. At the same time, the system must provide the possibility of extending and specializing the search for a given domain or a given application.

There are several approaches to achieve semantic interoperability when dealing with different ontologies. Wache [6] defined three ways for integration of ontologies by using single ontology, multiple ontologies or hybrid ontology approaches. It is intuitive that using just a single ontology is good for integration purposes, but it can overcomplicate the ontology and it is not very flexible under a cross-domain perspective. When using several ontologies there is a need of creating mappings between similar concepts in all the ontologies in order to achieve interoperability. This can become too complicated when dealing with several ontologies and hinder the introduction of new domains. The hybrid approach is based on using a common shared ontology and a set of local or application ontologies that are mapped uniquely to the shared vocabulary. In this sense the local ontologies would extend the vocabulary to the needs of a given domain and the interoperability is achieved based on the shared ontology. TaToo follows the hybrid approach and currently evaluates the most widely used ontologies in the environmental field as candidates to be the basis of the shared ontology.

In TaToo, there is also the need of describing in a uniform way environmental resources. We understand an environmental resource as a web resource (being a web page, a document, a model, a service, etc.), which is defined using an URI. In this sense, we are defining within the shared ontology a minimal environmental resource model as a part of the shared ontology that will contain the minimal set of cross-domain concepts and properties.

TaToo ontologies will be based on existing W3C standards, particularly RDF (http://www.w3.org/TR/REC-rdf-syntax/), RDFS (http://www.w3.org/TR/rdf-schema/) and OWL (http://www.w3.org/TR/owl-features/). In the ontology engineering process we will also try to reuse as much shared vocabularies as possible such as DC (http://dublincore.org/), FOAF (http://www.foaf-project.org/), SIOC (http://sioc-project.org/), etc. TaToo will not provide an ontology engineering tool, but it will rely on existing tools for ontology engineering such as the NeOn Toolkit (http://neon-toolkit.org/) or Protégé (http://protege.stanford.edu/). Within TaToo we

use the NeOn methodology to build ontology networks and semantic applications (http://www.neon-project.org/nw/NeOn_Book/).

4 The proposed architecture

TaToo aim is to provide a set of functionality mainly to search, discover and tag resources with metadata in order to improve the discovery through the information enrichment process described above. In order to offer such kind of functionality, the TaToo framework has to provide a set of system components (server side) providing the implementation of the functionality, and a set of user components (client side), letting the end users interacting with the core components through Graphical User Interfaces (or eventually APIs). From a high level view, the TaToo architecture is composed of a basic set of functional blocks or building blocks, which groups components with respect to the specific functionality they participate to realise. This high level group of functional blocks is presented in Fig. 2. The figure shows five different tiers[1] and the high level building blocks taking part.

The *Presentation tier* contains the User Components building block since this block provides all the interfaces required by the end user to access the system and take advantage of the provided functionality. These components interact with the *System Components* through a set of Web clients (stubs), which contact the Web server counterpart through operation invocation.

The *Service tier* contains the Web services implementing the entry point of the system part of the framework. These services provide functionality to the Web clients interacting with the TaToo Core Components contained in the *Business tier* (the *Business tier* implements the business logic of the entire framework). The Data tier is responsible of storing resource metadata (in form of RDF triples or structured) in a relational data base.

Finally, the *Cross tier* contains all the building blocks dealing with functionality generally known as cross-cutting. These building blocks provide functionality required by the framework to deal with requirements cross to all the previous tiers, such as security, in terms of authentication, authorization, access control, administration; and so forth.

Fig. 2 shows as well the different so far identified resources TaToo foresees to deal with. In general, Web services (SOAP or RESTful), and Catalogues are of primary interest. Other possible resources are models or Web pages containing structured or unstructured content. These resources have to be discovered, tagged, and evaluated. Evaluation and tagging of resources by the end user make possible the information enrichment process; searches will be more and more effective as each time based on a larger amount of available metadata. The high level architecture can be detailed as presented in Fig. 2.

[1] It is worth noticing that in the context of architectures based on the service-oriented paradigm, the term ‘tier’ is preferred to the more common term ‘layer’

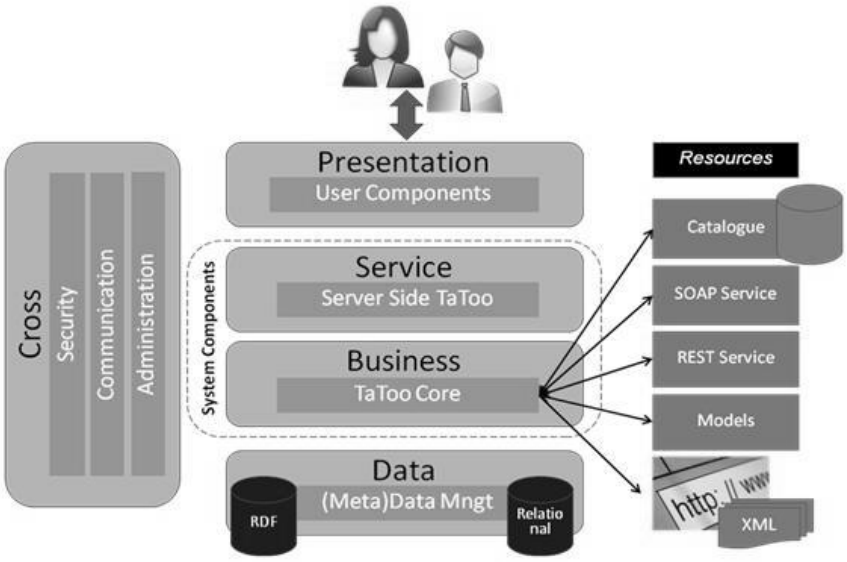

Fig. 2. TaToo High Level Architecture

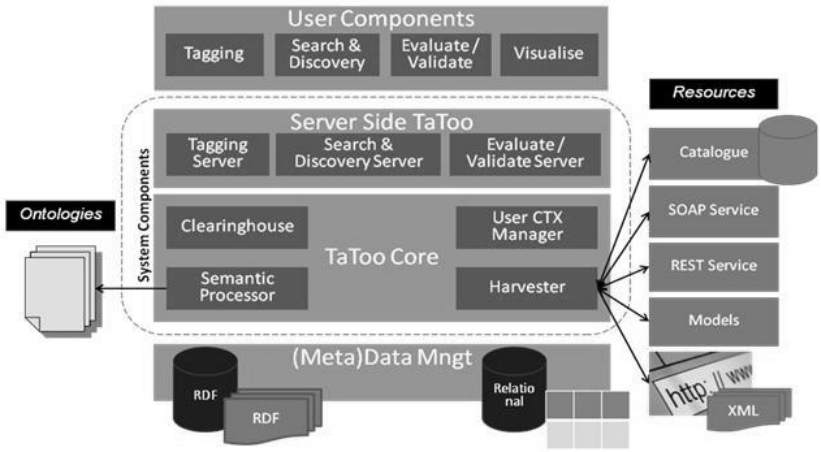

Fig. 3. TaToo Initial Architecture

The User Component building block contains four different sub-blocks. Tagging, Search & Discovery, and Evaluate / Validate offer the respective client side components to exploit system corresponding functionality. In general, these components are supposed to be directly used by the end user (this is particularly true in the case the component provides a GUI). The end user interacts with the component in order to take advantage of the TaToo offered functionality. In general, these components can be implemented as:

- Portlets making up a portal;
- External applications to be installed in the end user machine;
- Browser plug-ins to be installed in the end user browser;
- APIs in the case the end user wants to write his own custom application.

The general idea of TaToo is to realise a portal providing functionality through a set of configurable portlets (see Fig. 4).

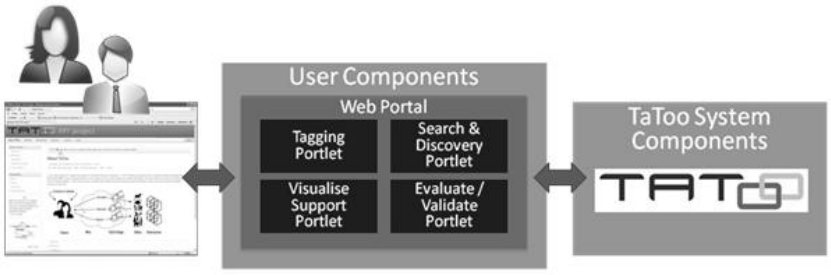

Fig. 4. TaToo Web Portal

The *Visualize* sub-block (still in the User Components block) contains tools for visualising resources. Once discovered, resources have to be accessed and visualised before tagging. Resources can be accessed through several means and protocols, such as HTTP(S); FTP(S); GridFTP; OpenLDAP; OGC WCS, WMS, WFS; or even Torrent (security and access control issues have obviously to be considered). Once the resource is available, it has to be visualised in order to be tagged. Graphics (Photos/Pictures etc.) in usual formats such as .jpg, .bmp, etc. or a .doc file can be easily visualised through applications that are normally available and installed client side. In the case of custom or proprietary formats, new tools for visualising resources are required. Visualising tools, even if not the main scope of the project, are fundamental as the tagging process can be performed by the end user only if they are able to 'visualise' i.e. to obtain an understandable representation of the resources being tagged.

The Server Side TaToo building blocks contain corresponding blocks at client side (apart from the block *Visualise* that has not correspondence at server side). As already stated, these are generally Web services offering a kind of interface to the underneath TaToo Core components.

The TaToo Core building block is the main part of the architecture and contains 'core' TaToo components implementing the business logic. In particular four components are of major importance:

- The clearinghouse plays the role of organising the semantic information on environmental resources. It is a central component for accessing the metadata storage and serves also as an information exchange support between the core system components;
- The Semantic Processor is the fundamental component dealing with Semantics. It uses a set of (pluggable) ontologies (in the environmental

domain) to provide functionality based on semantics. In general, it relies on an application framework (such as Jena) and a reasoner (such as Pellet, or the Jena embedded reasoner) to provide its functionality. An application framework provide useful APIs to manipulate RDF, support SPARQL query, and others;

- The User CTX Manager is in charge of managing the user context. In particular it offers the user the possibility to store information about the performed search, provided tags for resources, and others. For instance, it allows the user to retrieve passed performed search allowing the tagging process to be postponed in time (e.g. in the case the evaluation of the resources requires time to be performed[2]);
- The Harvester is the component capable of retrieving external resources (and associated metadata) that could be either data or associated metadata stored in catalogues, Web services or information contained in Web pages. The harvester plays the role of retrieving already available resource metadata (mainly from the resource owner). This means that in addition to the information enrichment process, metadata can also be collected harvesting available catalogues (or Web resources in general). The harvesting process can take place at system deployment time to create an initial set of metadata, and / or once in a while to get updated content.

Finally, the (Meta)Data Management building block deals with the storage where metadata on resources are stored (both resource owner metadata and metadata provided by end users through tagging). For proving TaToo's objectives and functionality three validations scenarios have been installed.

5 Validation scenarios

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 addressing the following environmental domains: climate change, agriculture, and anthropogenic impacts of pollution. Even if they have independent, distinct goals, they will be presented altogether in order to demonstrate complementarities of TaToo features, and improvements in the discovery process that are facilitated with TaToo. The three validation scenarios are further explained in the followings.

[2] Tagging can be performed as soon as the user evaluated the found resources. If this is trivial and immediate for a resource like a picture, it can be demanding and time consuming for collection of raw data, for which a long elaboration could be required

5.1 Climate Twins validation scenario

We call 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.

5.2 Agro-environmental validation scenario

In the agro-environmental scenario 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. That requires 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 alternative options for estimating/generating variables.

5.3 Anthropogenic impact of pollution validation scenario

The anthropogenic impact of pollution case study will enable the synthesis of existing (air) pollution monitoring databases, with epidemiological data 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 information description (metainformation) embedded into the appropriate semantic environment.

5.4 Overall project evaluation

Each validation scenario identifies several use-cases that demonstrate clear improvements in the discovery process of environmental resources. In TaToo project, we aim to track all the improvements, organize, and study them, in order to evaluate the overall impact achieved by the TaToo project. We intend to classify improvements against two major axes: type of improvement and type of beneficiaries.

We identify the following two types of improvements that will come as a result of TaToo tools usage:

1. *Improvements in the discovery performance*, i.e. more efficient discovery of resources. In this case, use-cases demonstrate that the adoption of advanced semantics and TaToo tools has improved the situation in measurable ways (i.e. by identifying certain indicators of performance that include result completeness, robustness, response time, or other appropriate measures to be defined in the use cases).
2. *Improvements in the discovery process*, i.e. new search and discovery experience not possible before. In this case, TaToo tools enable completely new discovery processes that have not been available with the traditional tools, or their usability are significantly improved with TaToo. For these cases quantitative measurement of improvement is not possible, rather a qualitative one is adopted. Very important in these cases is to justify existing tools cannot support such processes.

We also identify three major types of beneficiaries (user roles) for TaToo:

a. *Resource provider*, i.e. the one who supplies the resource, and in the domains we are treating could be a modeller, or a scientist.
b. *Resource publisher*, i.e. the one who is in charge of making the resource available. This user could coincide with the provider, but maybe not.
c. *Resource consumer*, which is the one who wants to discover the resource through some kind of querying, subscription to feeds or through a suggestion mechanism.

All use cases of the three validation scenarios will be classified against the two axes, in order to visualize them on the plane of beneficiaries and improvement types (see Fig 5). Certainly, all types of improvements are not relevant for all beneficiaries, and not all combinations are in the scope of the project. Core interests of TaToo remain in (a) process and performance benefits for the publisher (shown in dark gray), and (b) process improvements for the resource publisher and performance improvements for the resource consumer (shown in light gray).

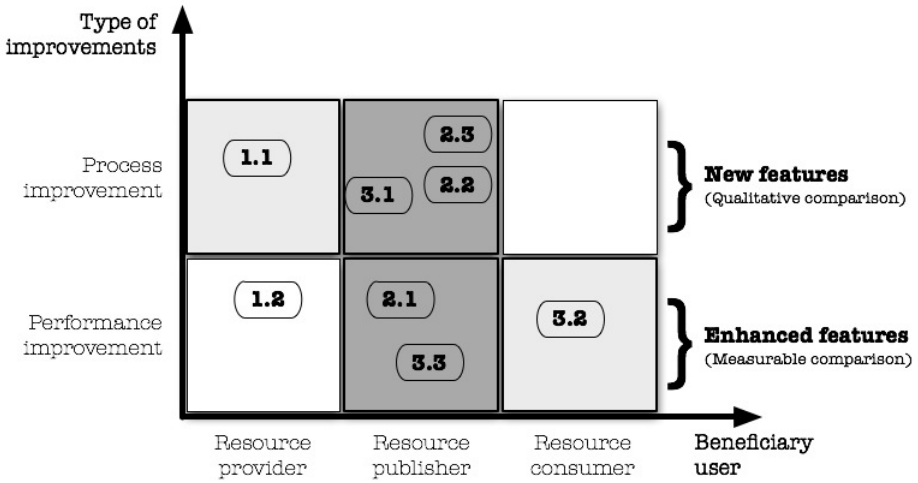

Fig. 5. The TaToo project Validation Matrix (The ids of the use-cases are examples)

6. Conclusions

Currently, we are at the beginning of the project and there will still be adjustments on the initial architecture and related requirements stemming from the validation scenarios, but what we clearly envisage and what you can read from the sections above is to mitigate the burden of meta-information providing by the system developers. Our aim is to do that in a community driven way, i.e. allowing different communities as well as individuals to add information to resources related to their view on them and embedded in their respective semantic environment.

At the end of the project, the TaToo users will benefit from having enriched their resources with semantics, embedded them in a semantic framework to share and reuse with their scientific community.

From the technical point of view TaToo will realise a framework able to cope with specific needs of the TaToo Validation Scenarios based on specific environmental domain ontologies. At the beginning the main technical focus will be on semantically supported tagging, search and discovery functionality. But, the long term vision is to provide a general framework able to support the entire environmental domain.

Acknowledgment: The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr. 247893T

References

1. Gordon, M., and P.Pathak, Finding information on the World Wide Web: the retrieval effectivness of search engines, Information Processing and Management, 35(2), 141-180, (1999).
2. A.E. Rizzoli and G. Schimak et al.:TaToo: Tagging environmental resources on the web by semantic annotations; Proceedings of 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; (2010).
3. 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).
4. Berners-Lee, T., Hendler, J. and Lassila, O.: The Semantic Web, Scientific American, (2001).
5. Gruber, T.: A translation approach to portable ontology specifications. Knowledge Acquisition 5, (1993).
6. Wache, H., Scholz, T., Stieghahn, H. and K'onig-Ries, B.: An integration method for the specification of rule–oriented mediators. Proceedings of DANTE'99, Kyoto, Japan, (1999).