An Architecture for the Semantic Enhancement of Environmental Resources

P. Dihé, S. Frysinger, R. Güttler, S. Schlobinski, L. Petronzio, R. Denzer, S. Nešić, T. Lobo, G. Schimak, J. Hřebíček, et al.

An Architecture for the Semantic Enhancement of Environmental Resources

P. Dihé,[1] S. Frysinger,[2,3] R. Güttler,[2] S. Schlobinski,[1] L. Petronzio,[4] R. Denzer,[2] S. Nesic,[5] T. Lobo,[6] G. Schimak,[7] J. Hrebicek,[8] M. Donatelli,[9]

[1]cismet GmbH, Altenkesseler Strasse 17 D2, 66115 Saarbrücken
[2]Environmental Informatics Group (EIG), Goebenstrasse 40, 66117 Saarbrücken, Germany
[3]James Madison University, Harrisonburg, Virginia, USA 22807
[4]Telespazio S.p.A, Via Tiburtina, 965 - 00156 Rome - Italy
[5]IDSIA - USI/SUPSI, Galleria 2, 6928 Manno, Switzerland
[6]Atos Research and Innovation, C/Albarracín 25, 28037 Madrid,Spain
[7]AIT Austrian Institute of Technology GmbH, A-2444 Seibersdorf, Austria
[8]Masaryk University, Kamenice 126/3, 625 00 Brno, Czech Republic
[9]Joint Research Centre of the European Commission, Ispra, Italy
pascal.dihe@cismet.de

Abstract. The vision of a Single Information Space in Europe for the Environment (SISE) requires seamless access to environmental resources, including data, models and services. Standardization organizations like OGC and OASIS have laid the foundations for interoperability on a syntactic level for many aspects of distributed environmental information systems (e.g. OGC SWE for sensor information). At the same time, the EC has undertaken a considerable effort to commit European stakeholders to offering their environmental information in such a way that it is accessible by interested parties, both on the scientific level by supporting research projects, like ORCHESTRA and SANY, and on the legal level by introducing directives (such as the INSPIRE directive). This development, amongst others, has led to the present situation in which a large number of environmental information sources are available. However, to implement the vision of the SISE it is not enough to publish resources. Environmental information must be discoverable, and it must be 'understandable' in different contexts in order to be used effectively by parties of various thematic domains. Therefore, in order to foster the implementation of SISE, semantic interoperability is a necessary element. Key to semantic interoperability is the presence of meta-information which describes the concepts of the environmental resources. Producing this meta-information puts a heavy technological burden on the individual resource providers such that it seems unlikely that enough semantic meta-information will ever be made available to reach semantic interoperability and thus accomplish the vision of SISE unless other ways to provide this essential meta-information are found. In this paper we introduce an architecture, developed in the FP7 project TaToo (247893), which tries to overcome the aforementioned obstacles by providing the possibility to easily annotate and rate environmental information resources, even by parties which do not own the resource, and

transparently equipping this information with domain knowledge and thus enhancing discoverability and usability of resources with semantic technologies. The objective of the architecture is to seamlessly blend in with existing infrastructures by making use of *de facto* standards while offering support for discovery, annotation and validation of environmental resources through open interfaces.

Keywords: SISE, semantics, search, discovery, annotation, tagging, resources,

1   Introduction

Search and discovery have been staples of the modern Internet culture since the release of the first practical browser. But these were followed in short order by the need to save discovered information through "bookmarking." This allowed an individual user to keep track of web-based resources that were interesting to them, and it added value to the information in that the facts of its prior discovery and selection constituted some additional information: acceptance. That is, the individual had already searched for, discovered, and somehow approved of the resource, and creating a bookmark captured this additional information.

But is that the limit of the possibilities of web-based information enrichment? Could a user attach more detailed semantic content to a discovered resource to help them use it in the future? Could the attachment of such information be independent of the particular computer on which they were working? Could they share such semantic enhancements with others? Would such capabilities facilitate the use of highly distributed web-based resources – such as environmental data and services – in the solution of complex problems being addressed by many individuals and organizations distributed across many organizations and nations?

These questions are critical to the notion of a Single Information Space in Europe for the Environment (SISE), an important effort on the part of the European Commission to harmonize and facilitate environmental decision-making. The vision of SISE requires seamless access to environmental resources, including data, models and services. But these resources themselves also have the potential for semantic content that is related to the viewer of the resource.

Consider a simple example. A restaurant aficionado might find a reference to a particular Italian restaurant and, having been pleased by the chef's generous use of garlic, might annotate this restaurant favourably. This will be useful when this particular customer is choosing a restaurant for a future dinner. But it might also be useful for another user who, sensitive to garlic, would choose to avoid this restaurant because of this very same characteristic.

One could imagine a very similar situation occurring within the context of environmental data and services. A particular data set, for example, may be valuable for some applications because of a characteristic that, for other applications, renders it inappropriate.

Environmental data and models are generally very dependent on the conditions of their capture or generation, respectively. Rarely would an astute environmental

manager base a decision on a data set without understanding the conditions under which the data were obtained, in order to ensure that those conditions were consonant with the decision problem at hand. Similarly, the use of modelling results must be conditioned on the consistency between the assumptions made in the model with those in the question being decided.

In order to achieve the objectives of SISE, it will be necessary to find a way to capture and report "auxiliary semantics" associated with data and services discoverable on the web. These auxiliary semantics would, in general, be independent of the original author of the data or service, and would be understood to be subjective, at least to the extent that they represent the interpretation of a viewer of the resource, as opposed to the generator of the resource.

2   Requirements of SISE

As described by Hřebíček and Pillmann [1] and the Inspire Forum [2], the concept of the **Single Information Space in Europe for the Environment (SISE) was introduced in the European Commission's Framework Programme 7 (FP7) in association with objective** ICT-2007.6.3: ICT for Environmental Management and Energy Efficiency. The central idea is to support real-time connectivity to and seamless search of multiple environmental resources, across borders and disciplines, with facilitated data acquisition and sharing, along with service chaining on the web. Several workshops were conducted.

As reported by O'Flaherty [3], following an Expert Consultation Workshop in Brussels in February of 2008, "[t]he central requirement of the SISE will be to enable technical interoperability, driven by two key trends:

1. Complexity Management: requiring a holistic approach, multi-disciplinary research, and an infrastructure accessible to all disciplines, communities and actors.
2. Environmental Legislation in Europe: driving higher requirements for common environmental monitoring and reporting."

Six dimensions of SISE were identified at this workshop:

1. User Dimension: Actor centric and adaptable to the expectations and needs of many types of users.
2. Content Dimension: A global information space including models, knowledge, services and tools.
3. Data Dimension: Fusing multiple heterogeneous sources, sensors, platforms, networks and visualisations.
4. Modelling & Decision Support Dimension: Involving distributed virtual teams and communities
5. Information and Service Dimension: Stable and secure multilingual services, using semantically enhanced, *ad hoc* on-demand service chaining on the web.
6. Other SISE Dimensions: Including standardisation, ethics/privacy, data policy, IPR and liability

The assembled experts agreed about the following prioritisation of topics for future research and innovation actions on the SISE:

These dimensions and priorities lead naturally to some high-level requirements of SISE, which may be briefly summarised as follows:

- Support the discovery of relevant resources
- Facilitate integration of and seamless access to data sources residing on a standard based infrastructure (such as those provided by SANY or ORCHESTRA)
- Support repositories (e.g. databases, caches, inventories) for quality controlled and securely managed resources and their results
- Incorporate integrated security in order to control access to resources (e.g. data sources, catalogues, whether one's own or third party resources)
- Preserve the possibility to publish results on the web for public access
- Provide easy to use tools and user-friendly services and interfaces, e.g. access control, workflow management, delivery management, visualisation, data extraction, and administration, embedded in the users' semantic context.

The deployment of a Single Information Space for the Environment in Europe is inhibited by the lack of a mechanism allowing data and service users to provide annotations that would add value to these resources. A middleware infrastructure is needed to fill this gap between environmental resources and end users. This framework needs to facilitate the life-cycle utility of environmental information from its collection and persistent storage to its discovery and purpose-oriented exploitation. Such a goal can be achieved through coherent, transparent, efficient and context dependent (e.g. semantically enhanced) discovery mechanisms, pre-processing services that translate meaning and structure of information, accompanied by services or information items (e.g. data) tailored to end-user needs, and explanatory facilities.

TaToo's open approach to this need allows it to contribute not only to a single European Information Space for Europe, but also to the requirements of sharing information as expressed in the requirements of the SEIS (Shared Environmental Information System).

3 Needs, Gaps & Objectives of TaToo

Given the objectives and requirements of SISE, as described in Section 2, the TaToo project targets one major objective:

To contribute to the elimination of obstacles in discovery as well as the context-sensitive interpretation and use of environmental resources within SISE by developing easy to use tools within a semantic framework for discovery, access, and interpretation in a multilingual and multi-domain context.

This objective was approached first by conducting two needs & gaps analyses to identify the most pressing issues regarding the enhancement of the annotation and discovery process of environmental resources (i.e. data and services). Users were asked which types of environmental resources are important to them (e.g. geospatial data, structured raw data, time series, etc.); for which purposes the user needs the

discovered resources (e.g. input data for models); how the user currently searches for relevant resources (e.g. search engines, data catalogues, etc.); and how discovered resources can currently be accessed (downloads, web services, etc.).

Users are mostly interested in various types of data ranging from time series data to data aggregated in reports or documents, as well as software and mathematical models. Common to most types of data requested is that they should be related to a specific geographical location. Users particularly want additional meta-information that helps them to assess whether a resource is usable for their clearly defined purpose. This includes, for example,

- information about the quality of the data,
- how data were collected, and by whom,
- whether and which interpolation methods were applied, and
- the applicability of models and algorithms to a certain domain.

Thus, one of the crucial features of TaToo is the ability to provide appropriate and rich semantic meta-information about a resource during the discovery process.

Regarding the discovery of environmental resources, most users rely on classical search engines but would prefer a solution that combines the usability and simplicity of search engines with the power and advanced features of semantic discovery. The most common problems that need to be addressed are related to low precision of the results, because the search results are often cluttered with too many irrelevant or unsuitable resources. Another issue is that resources exposed through services (in particular OGC services) often cannot be found by classical search engines. Users would also like to be able to perform discovery that takes their context (e.g. user profiles, user relationships, user location, search history, etc.) into account in order to improve results or to discover new relevant resources. Furthermore, they want to be able to search for links between related topics, e.g. for data that are available for the same spatio-temporal scope and pertain to two different thematic domains (e.g. distribution of persistent organic pollutants and cancer incidence statistics).

Regarding the different means to tag resources, or to view, alter and evaluate tags attached with resources, users raised the concern that multiple tags for the same meaning could emerge. They recommended that TaToo take standardised thesauri and dictionaries into account and provide a mechanism to propose new tags that can be added to existing ontologies.

Most users would use a portal as the main system entry point, but integration of TaToo functionality with their current systems is also a topic of great interest. Therefore, they appreciated the possibility of accessing the TaToo framework through public service interfaces in order to create their own custom clients.

Concerning the question of whether and how the users would be willing to use and contribute to TaToo, the vast majority indicated that they would like to use TaToo to search for and interpret environmental resources. Users also requested tools to enable them to tag resources in a manner specifically related to their field of expertise in order to improve the discovery of environmental resources in general. Furthermore, the ability to easily register their own resources and to harvest an initial set of meta-information during a (semi-) automatic registration process is a highly desired feature.

4   The TaToo Architecture

TaToo's primary aim is to provide the functionality to discover and tag resources with metadata in order to improve future discovery through the information enrichment process described above. In order to offer this kind of functionality, the TaToo framework has to include a set of system components (on the server side) providing the implementation of the functionality, and a set of user components (on the client side) helping the end users interact with the core components through graphical user interfaces (GUI) [4].

In the context of the TaToo objectives stated previously, the TaToo architecture must incorporate a number of properties and principles, such as the re-use of existing software wherever possible, in order to avoid wasted development effort, and the use of data- and interface-related standards from the relevant communities, where appropriate, to assure seamless integration into existing infrastructures. Moreover, TaToo must provide easy to use GUI clients for all TaToo functions to a wide range of users, and provide well-defined interfaces (e.g. APIs and standardised web service interfaces) allowing third parties to embed TaToo functionality in their own applications.
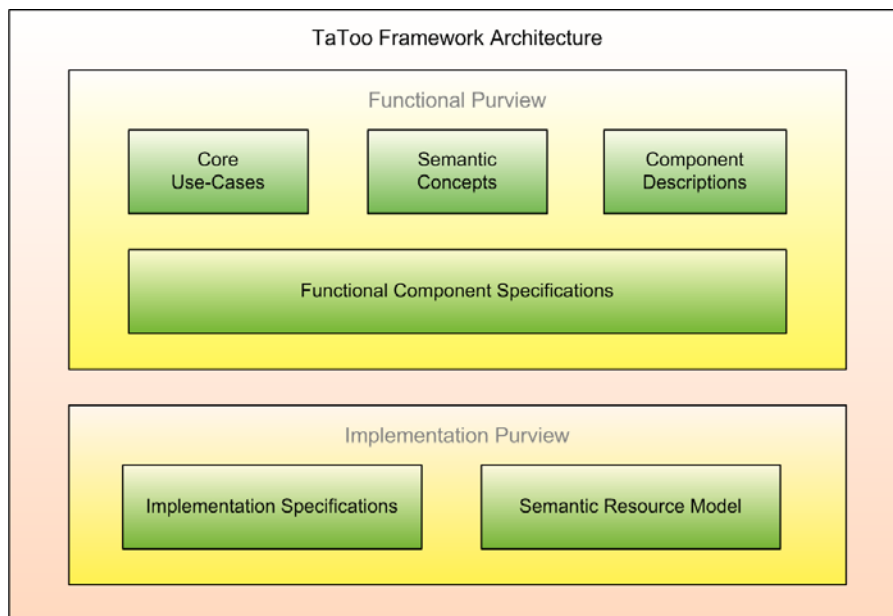
The TaToo Framework Architecture has to consider the facilities for data and information discovery, visualisation, and evaluation/validation, as well as tagging/annotation by humans and the harvesting of meta-information about resources by (semi-) automated processes.

A coherent architectural design approach, especially in the context of SISE and the TaToo objectives stated in Section 2, demands careful consideration of crucial functional requirements as well as the adoption of well-proven architectural design procedures which result in service oriented architectures that meet essential architectural properties like openness, interoperability, extensibility, flexibility and loose coupling. For this reason, the conceptual foundations of the TaToo architecture follow an evolutionary approach based on experience from former infrastructure projects and sound procedures already established in ORCHESTRA [5] and SANY [6]. Among the predominant procedures which are applied in the architectural design and specification of the TaToo framework are the separation of concerns, the logical grouping of functionalities, the specification of open and standards based interfaces, the introduction of an implementation independent layer, and the establishment of a domain-independent generic infrastructure. This ultimately leads to a multi-tier architecture composed of interacting components, which is not influenced by implementation specific restrictions, technological choices, or a specific thematic domain.

When talking about conceptual foundations of the architecture, one has to also consider supporting elements like guidelines, templates and tools that help architects and developers to specify and develop their components and domain ontologies in a structured, methodical and consistent manner. In this spirit, the TaToo architecture doesn't only specify the TaToo framework, but it also provides both the methodology and the tools to effectively facilitate the initial architectural specification as well as the further extensibility of the TaToo framework, including the support for arbitrary

thematic domains. Extensibility is a key feature for TaToo, since it allows third parties to register their own resources for semantic annotation, to plug in their own domain ontologies and to contribute to the TaToo framework itself with their own specialised client applications and semantic tagging and discovery services. Thus, the TaToo community will greatly benefit from the framework's open nature.

The TaToo architecture consists of an implementation independent Functional Purview and an Implementation Purview (Figure 1). Implementation independence at the architecture level leads to a more sustainable architecture that is not vulnerable to technology changes in the future, and it is thus able to accommodate changes in technology without changing the architecture itself. This is especially important because TaToo intends to make use of existing solutions, such as widely adopted semantic frameworks like Sesame and Jena.

Fig. 1. TaToo Framework Architecture

The Functional Purview therefore introduces an implementation independent layer that is mapped to a concrete implementation platform, resulting in a reference implementation of the TaToo framework. At first, the Functional Purview specifies the conceptual foundations of the architecture, including the definition of the architectural design process followed by its objectives and structural elements (Tiers and Building Blocks). It also explains the considerations and decisions that resulted in the chosen architectural definition process, which follows widely accepted principles of software architecture. Additionally, it establishes the conceptual principles for discovery and tagging in TaToo as well as an ontology framework that allows the production of formal, semantically enhanced resource descriptions. Finally, it provides the actual implementation independent specification of the TaToo

framework consisting of the descriptions and functional specifications of all TaToo components, and an illustration of their basic interactions with the help of a set of core use-cases.

The Implementation Purview complements the Functional Purview in the areas of implementation and technology. The focus of the Implementation Purview lies on how a certain component is realised in detail, the formal specification of the component's interfaces and information exchanged, and the specification of a Minimum Environmental Resource Model (MERM).

The elements of architectural design that define the overall structure of the TaToo Framework are Tiers and Building Blocks. Tiers logically separate general concerns, like presentation and business logic, while Building Blocks and the components defined therein share a common set of properties and organise and group more concrete functionalities.

The specification process in service-oriented architectures often stops at the service interface level. The TaToo architecture goes one step further by specifying the functionalities of client components and underlying business logic components in addition to the publicly visible functionality of the system exposed through the interfaces and operations of services.

Therefore, Figure 2, which shows only Building Blocks, illustrates some essential components and the most important information flows. The TaToo framework architecture is designed as an n-tier architecture that currently consists of the following 4 tiers:

1. The Presentation Tier, which is concerned with user interaction and the presentation and aggregation of information.
2. The Service Tier decoupling the Business Tier and Presentation Tier, as well as serving as a layer to enforce interoperability through the provision of well defined and self-describing service interfaces.
3. The Business Tier, which is responsible for the core functionality (business logic) of the TaToo System.
4. The Data Tier, which is concerned with the storage of semantically enriched information, and other data (registered resources to be harvested, user's information, etc.).
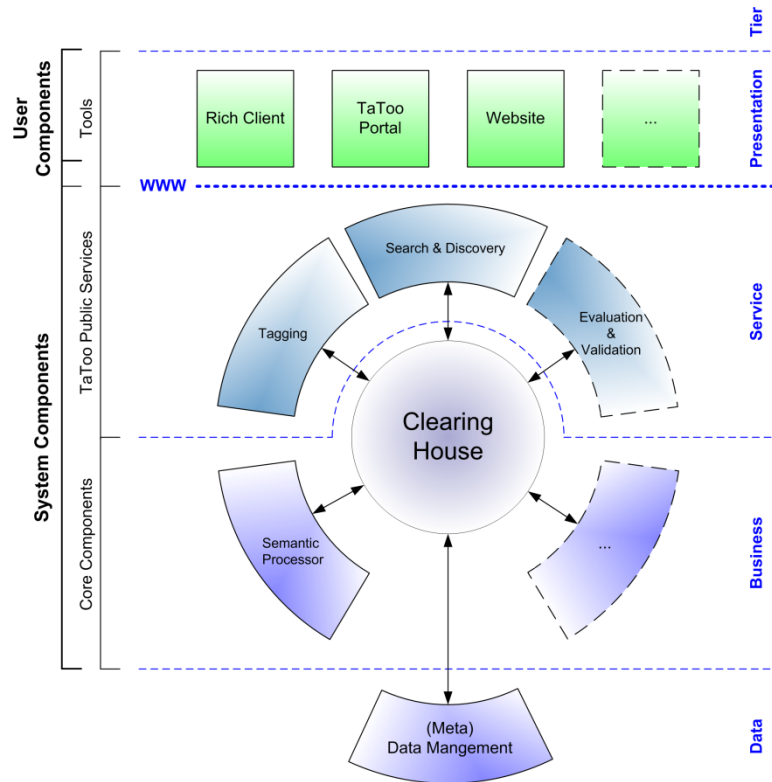
Fig. 2. TaToo Framework Architecture Tiers and Building Blocks

Besides the tiers, the second elements of architectural design are the Building Blocks (Figure 3). The architecture is composed of two main high level Building Blocks: the User Components Building Block and the System Components Building Block. The User Components Building Block is composed of User Components (also called TaToo Tools), such as tagging and discovery tools, with a graphical user interface that supports the interaction with a human user.
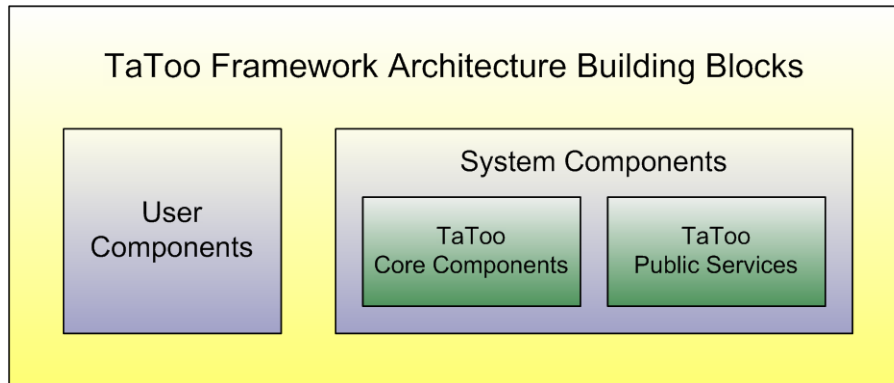
Fig. 3. TaToo Framework Architecture Building Blocks

The System Components Building Block is further decomposed into the Public Services Building Block and the Core Components Building Block. The Public Services Building Block is composed of Public Services, like tagging and discovery services, which expose TaToo functionality to User Components through well defined (specified) interfaces. The Core Components Building Block provides the business logic, data and meta-information management, thus insulating complex semantic procedures from the rest of the system. These components are only accessed by other System Components and by the TaToo Public Services. Consequently, interoperability is less of a concern and the System Components may, but need not, be accessible over standardised web service interface.

5   TaToo Architecture – Added Value for SISE

TaToo is very ambitious in its goal of providing added value to the domain of discovery and annotation of environmental resources. TaToo provides the appropriate means to create easy to use discovery mechanisms and tools, and facilitates discovery with each iterative tagging cycle. Discovery becomes more precise and information quality and value are increased.

TaToo mitigates the burden of creating 'meaningful' - i.e. semantically enhanced - meta-information by preparing the ground for information enrichment of the discovered resources. One of TaToo's benefits is that when handling a resource with a known semantic context TaToo supports users in their effort to add annotations and ranking information (e.g. according to importance, quality, uncertainty, etc.).

More concretely, TaToo delivers tools facilitating discovery of environmental resources through the use of underlying semantic structures. It provides open specifications of semantic discovery frameworks and tools and substantially contributes to the fulfilment of the requirements of SISE. TaToo provides a public Service tier to deal with arbitrary resources and in particular with resources that are based on open standards. TaToo's ability to deal with a variety of ontologies in a

multilingual context, depending on the current discovery strategy, strives to overcome well-known problems in this area.

Due to the iterative improvement and enhancement of descriptive information through TaToo's tagging cycles, supported by an intelligent tagging process, decision support quality increases significantly.

Going back to the SISE imperatives identified in Section 2, we can explicitly state the manner in which TaToo supports fulfilment of each of these imperatives.

Regarding the six dimensions of SISE:

1. User Dimension: Actor centric and adaptable to the expectations and needs of many types of users.

The TaToo architecture deals with this dimension on technical as well as functional levels. On the technical level the TaToo architecture foresees user components including a portal that can be tailored to different user types, and specific user-centric client applications that can use the public service interface of the public services (such as tagging and discovery). On the functional level the TaToo architecture establishes no restrictions on the type of resources for which TaToo functionality is offered, thus supporting the needs of many types of users.

2. Content Dimension: A global information space including models, knowledge, services and tools.

The TaToo architecture is open to easy integration of new services, tools, knowledge, or any kind of resources.

3. Data Dimension: Fusing multiple heterogeneous sources, sensors, platforms, networks and visualisations.

Any kind of resource coming from any kind of source can be the object of TaToo functionality (tagging, discovery, evaluation, etc.).

4. Modelling & Decision Support Dimension: Involving distributed virtual teams and communities

TaToo supports decision-making by offering the decision maker valuable meta-information of different kinds (e.g. "rating information", evaluation information about resources of different types) introduced by other users, thus supporting community building.

5. Information and Service Dimension: Stable and secure multilingual services, using semantically enhanced, *ad hoc* on-demand service chaining on the web.

The TaToo architecture anticipates a set of public services designed according to standards or best practises with standardized security measures. Semantic meta-information supports the use of these services.

6. Other SISE Dimensions: Including standardisation, ethics/privacy, data policy, IPR and liability

The TaToo architecture calls for the use of standard approaches wherever possible.

Regarding the technical requirements of SISE:

- Support the discovery of relevant resources

The TaToo architecture offers a public discovery service making use of different kind of semantic meta-information.

- Facilitate integration of and seamless access to data sources residing on a standard based infrastructure (such as those provided by SANY or ORCHESTRA)

TaToo offers meta-information about resources like data sources, including information about access procedures and semantic meta-information about suitability of the use of a data source in a given situation.

- Support repositories (e.g. databases, caches, inventories) for quality controlled and securely managed resources and their results

TaToo doesn't directly support repositories, but it supports the discoverability and usability of repositories by offering corresponding meta-information.

- Incorporate integrated security in order to control access to resources (e.g. data sources, catalogues, whether one's own or third party resources)

The TaToo architecture provides security measures for access by the public to meta-information about resources and TaToo services, but TaToo does not directly support security concerning the resources themselves. Nevertheless, meta-information about a resource offered by TaToo can include meta-information concerning access control for a resource.

- Preserve the possibility to publish results on the web for public access

The TaToo architecture supports the discoverability of published results as well as meta-information about them (e.g. concerning their usability for a given purpose).

- Provide easy to use tools and user-friendly services and interfaces, e.g. access control, workflow management, delivery management, visualisation, data extraction, and administration, embedded in the users' semantic context.

The TaToo architecture foresees different kinds of user components offering user-friendly access to TaToo functionality, such as a portal, or public services to be accessed by user-centric client applications.

6   Generalization and future research

In this paper we have introduced an architecture for the semantic enhancement of web-based environmental resources that serves to enhance the discoverability and usability of these resources with semantic technologies in the context of the vision of a Single Information Space in Europe for the Environment (SISE). The architecture is built upon widely accepted architectural design principles and addresses in its first

iteration the core functionalities, tools and concepts necessary to leverage semantic discovery and annotation of resources in a multi-domain context.

Future work on this architecture will take into account further research and innovation on SISE and will be engaged in activities to

- continuously enhance existing services and tools and identify new services and tools according to user needs
- develop an integrated user context management and access control concept incorporating the Friend of a Friend FOAF vocabulary and the OpenID standard for decentralized authentication.
- update TaToo's environmental resource model and the adjacent ontology framework in order to leverage domain ontologies development and provide support for cross-domain mappings
- invent new discovery strategies to improve context-aware and domain-oriented discovery
- develop a common model that allows inference with ontologies in different languages, in order to support multilingual tagging and discovery
- adopt the so-called linked data approach for the semantic web [7] in order to enable TaToo to join the ever-growing linked data community, and thus to access and discover semantically annotated resources on a global scale

Acknowledgments

References

1. Hřebíček, J. and W. Pillmann, Shared Environmental Information System and Single Information Space in Europe for the Environment: Antipodes or Associates?, European conference of the Czech Presidency of the Council of the EU TOWARDS eENVIRONMENT, J. Hřebíček, J. Hradec, E. Pelikán, O. Mírovský, W. Pillmann, I.Holoubek, T. Bandholtz (eds.) Masaryk University, Brno (2009)
2. Inspire Forum, Single Information Space in Europe for the Environment (SISE), http://inspire-forum.jrc.ec.europa.eu/pg/groups/10035/single-information-space-in-europe-for-environment-sise/, (April 4, 2011)
3. O'Flaherty, J. J., Towards a Single Information Space for the Environment in Europe. Experts Consultation Workshop – ICT for Sustainable Growth – Brussels 15 Feb 2008. ftp://ftp.cordis.europa.eu/pub/fp7/ict/docs/sustainable-growth/sise-workshop-report-08_en.pdf (2008).
4. Schimak, G., Rizzoli, A.E., Avellino, G., Pariente Lobo, T., Fuentes, J.M., Athanasiadis, I. N. (2010), Information enrichment using TaToo's semantic framework; Proceedings of Metadata and Semantics Research (MTSR) Conference 2010, Alcala de Henares, Spain, S. Sanchez-Alonso, I. N. Athanasiadis, E. García-Barriocanal, N. Palavitsinis, http://www.ieru.org/org/mtsr2010/; (2010)

14

5. Usländer, T. (Ed.), Reference Model for the ORCHESTRA Architecture v.2.1, OGC "Best Practices" paper No. 07-097, http://www.opengeospatial.org/standards/bp, (2007)
6. Usländer, T. (Ed.), Specification of the Sensor Service Architecture v3.2, OGC discussion paper No. 09-132r1, http://sany-ip.eu/filemanager/active?fid=333, (November 15, 2009)
7. Bizer C., Heath T., Berners-Lee T., Linked Data - The Story So Far, International Journal on Semantic Web and Information Systems, Special Issue on Linked Data, http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf, (2009)