Towards a Semantically Unified Environmental Information Space

Saša Nešić[1], Andrea Emilio Rizzoli[1], and Ioannis N. Athanasiadis[2]

[1]IDSIA, Manno, Switzerland
{sasa,andrea}@idsia.ch
[2]Democritus University of Thrace, Xanthi, Greece
ioannis@athanasiadis.info

Abstract. In recent years we have witnessed a proliferation of environmental information on the Web thanks to advances in automated data acquisition and to the widespread use of computer based models and decision support systems processing environmental data. The number of environmental data providers has been also increasing. However, each provider manages its own data sets encoded into specific data formats and unaware of related and relevant data managed by other providers. Also, most of the environmental data providers store their data into huge, centralized repositories, which makes the access and discovery of desired data difficult. The Linked Data principles along with the Semantic Web technologies have been recognized as a promising solution to both environmental data integration and discovery. Unique identification of environmental data by HTTP dereferencable URIs, semantic annotation of environmental data by shared domain conceptualizations (ontologies), and interlinking of related environmental data by typed (semantic) links will enable the integration of disconnected environmental data sets into the semantically unified environmental information space. Semantic annotations and semantic links will then enable semantic discovery of environmental data over such unified information space. In this paper, we try to identify a number of requirements that environmental data providers should satisfy in order to make their data fully contribute to this vision. In particular, we are focused on requirements regarding environmental data identification, representation, annotation and linking.

Keywords: environmental data identification, semantic annotation, semantic linking.

1 Introduction

Over recent years, the adoption of the Linked Data best practices for publishing and collecting structured data on the Web [6] has opened the possibility of creating a unified information space, connecting data from different sources and domains such as weather forecasts, music stores, television and radio programs, on-line communities and business records. This information space is commonly

refereed as a Linked Open Data (LOD) Cloud and is considered as an incubator for the envisioned Web of Data. The main idea of the Web of Data is linking data instead of linking documents, which should enabling fine-grained integration of cross-domain information into a globally unified information space [3]. Moreover, the Web of Data has also been recognized as a foundation for the Semantic Web, which in spite of a number of different interpretations, has been recognized as a global Web of machine-readable data. Humans are the current Web's semantic component. They are required to process the information available on the Web to ultimately determine their meaning and relevance for the task at hand. The Semantic Web intends to move some of that processing to software agents [7]. In order to discover and map data more precisely, software agents require machine-readable data and machine-understandable data semantics (metadata). What the Semantic Web brings to the situation are the new data representation model (the predicate-based structures to express meaningful assertions) and the ontologies and rules to enable intelligent software agents to parse meaning from these assertions (sentences). Intelligent software agents will not be able to 'think' like their human counterparts, but they will be able to reason logically around the encoded explicit assertions, infer new ones, and assist humans in committing their tasks.

The Linked Data and the Semantic Web principles are universal; they are not restricted to any particular domain. As such, they represent promising solution for semantic integration of currently disconnected environmental data sets present on the Web. Traditionally, environmental data has been published on the Web as chunks of digital content, more frequently as text files, in some cases either stored as XML or marked up as HTML tables. Some HTML documents containing related environmental data are interlinked but the meaning of the relationships between the linked documents can only be implicitly distinguished. Hyperlinks indicate that two documents are related in some way, but it mostly left up to the human user to infer the nature of the relationship. HTML initially did not provide neither elements enabling typed links between documents nor between individual entities described in particular documents. Advances towards this direction, as microformats[1] has not been widely adopted either. Environmental data are no exception to this situation, while complexity, spatiotemporal reference, and uncertainty, make things even worse. Common practice has proven that environmental data are usually stored in non-reusable raw formats, situated in sparse locations and managed by different authorities, which ultimately raise obstacles in making environmental information accessible [1]. As a result of that, environmental data published on the Web looks like sets of disconnected data islands that are unaware of each other. Having environmental data published in accordance with the Linked Data and the Semantic Web principles would enable building of the *semantically unified environmental information space*, where environmental information becomes a common asset that is shared among peers, instead of a resource in scarcity that peers strive for [1].

---

[1] http://microformats.org/about

In this paper we discuss and analyze a set of requirements/principles that environmental data providers should respect in order to publish their data to the semantically unified environmental information space. We start the discussion by providing a brief overview of the use of the Semantic Web technologies in the environmental domain so far (Section 2). Then, we discuss and analyze requirements for environmental data identification and representation (Section 3). After that we continue with requirements for semantic annotation (Section 4) and linking (Section 5) of environmental data. Finally, in Section 6 we discuss and analyze requirements for semantic discovery (i.e., search and navigation) of the environmental data that are part of the unified environmental information space. Section 7 concludes the paper and outlines future work.

2 Application of the Semantic Web Technologies in the Environmental Domain

Environmental modeling and software are challenged to deal with complexity, uncertainty, scaling and integration issues, qualities inherited from the physical world, and thus comprise a challenging testbed for Semantic Web technologies. Environmental software embody sophisticated statements of environmental knowledge. Yet, the knowledge they incorporate is rarely self-contained enough for them to be understood and used by humans or machines without the modelers mediation [18]. Research efforts, so far has concentrated in three lines of work:

- **Semantics for environmental data annotation** through the development of domain vocabularies, thesauri and ontologies. There are several ongoing efforts on defining standards for sharing environmental data, and contribute to the Semantic Web vision, including these published by the US Environmental Data Standards Council (EDSC 2006), the standards developed gradually since 1994 by the European Environment Information and Observation Network[2] and the guidelines on vegetation plots and classifications of the Ecological Society of America (VEGBANK[3]). Also, the Ecological Metadata Language (EML)[4] provides with a metadata specification for describing data relevant to ecology. Food and Agriculture Organization (FAO) of the United Nations has made its thesaurus of food and agricultural terms, publicly available through the AGROVOC web services[5]. Towards the same direction contributes the OpenGIS specifications by the Open Geospatial Consortium[6] for the standardization of geo-referenced data, which are very common in environmental applications. The above are selected only as an indication of the parallel efforts for organizing and naming environmental

---

[2] `http://www.eionet.europa.eu/`
[3] `http://www.vegbank.org`
[4] `http://knb.ecoinformatics.org/software/eml/`
[5] `http://www.fao.org/agrovoc/`
[6] `http://www.opengeospatial.org`

data units using standard vocabularies and URIs for sharing them on the Semantic Web.

- **Semantics for environmental software annotation**: Environmental Modeling Frameworks typically offer though an API a collection of technical solutions for environmental scientists to program and deploy their models, and include facilities for data management and visualization, mathematical integration and scaling across time and space, among others. However, Environmental Modeling Frameworks are often invasive and heavyweight [11], thus software reuse comes with a high price. Integrating rich semantics in environmental model interfaces has a great potential to both maximize interoperability [16] and ensure sound integration [2]. For example, in the SEAMLESS project domain ontologies were adopted for the annotation of model interfaces, and used them for generating software code to facilitate software integration through a semantic-rich development methodology [1, 10, 9]

- **Semantically-aware environmental modelling** is a way of designing, implementing and deploying environmental datasets and models based on the independent, standardized formalization of the underlying environmental science [18]. It can be seen as the result of merging the rationale of declarative modelling with modern knowledge representation theory, through the mediation of the integrative vision of a Semantic Web. In this knowledge-driven approach, where the knowledge is the key not only to integration, but also to overcoming scale and paradigm differences and to novel potentials for model design and automated knowledge discovery [18]. Despite the clear potential offered by semantic modeling applied to environmental problems, only limited case studies are available, i.e IMA [17], ESD [19], SEEK [12], and ARIES [18]. The feasibility of wide adoption of the approach remains to be seen in the coming years.

While the visions of the Semantic Web and Linked Open Data have not come into life yet, there is a growing interest from the Environmental community on the field, and the semantically unified environmental information space is an era to come. In the following sections we preview how this will be achieved.

3   Environmental Data Identification and Representation

Before considering how environmental data should be identified we first need to define what kind of data is meaningful to be identified. We distinguish between the following concepts: *content unit*, *information unit* and *data unit*. Content units are units of raw digital content (e.g., text, graphics, audio, and video). Information units are pieces of information that can be understood when provided as standalone. Data units are content units which have one of more information units encoded inside themselves. In the envisioned semantically unified environmental information space, environmental data will be organized in data units, each one of which being characterized by a set of machine-processable descriptions and a set of relationships. The relationships can be established among

related environmental data units but also between them and other web entities as well as real-world entities and abstract concepts.

Linked Data vision builds directly on the Web architecture [8], thus environmental data units should be identified by unique resource identifiers (URIs). There has been a lot of confusion about URIs and URLs (Uniform Resource Locators) so far. The two concepts share the same meaning up to some extent, but are not equivalent. URIs identify resources either by location, or a name, or both. Accordingly, they can be classified as locators, or as names, or as both. A URI that identifies a resource by name in a given namespace, without defining how the resource could be obtained, is called unique resource name or URN. Unlike a URI, a URL identifies a resource by specifying a network location from where the resource can be obtained. Linked Data principles suggest using only HTTP URIs and avoiding other URI schemes such as URNs [13] and DOIs [15]. Two essential reasons for this are that HTTP URIs provide a simple way to create globally unique names in a decentralized fashion and that HTTP URIs serve not just as a name but also as a means of accessing the identified entity. In order to achieve not only global uniqueness but also good arrangement of the HTTP URIs within the environmental information space it is recommended that environmental data providers agree on using a standardized HTTP URI schema. One proposal for such an HTTP URI schema for the environmental could be as follows:

```
http://[domain name]/[provider OpenID]/[resource ID]
```

The parts of the schema are:

- **domain name** - a reserved DNS name for the environmental domain, it might be a sponsored top level domain .envi such as .mobi or .edu;
- **provider OpenID** - an OpenID identifier identifying the environmental data provider; OpenID is an open, decentralized standard for the authentication of online users;
- **resource ID** - a local resource identifier which is unique among the data units published by the data provider;

Having environmental data units uniquely identified by HTTP dereferencable URIs, the next principle that the environmental data providers should adhere to in order to make their data be seamlessly integrated to the semantically unified environmental information space is a universal data representation model. On the Web of Data, data is represented by the Resource Description Framework (RDF) data model. RDF is the data representation model especially designed towards the Web architecture and aiming at providing integrated representation of information that originates from multiple sources. Any resource published on the Web of Data is represented by unique RDF node and described by a number of RDF triples (sentences of a basic structure): *subject*, *predicate*, and *object*. The *subject* is the URI of the resource, the *predicate* indicates the relationship between the *subject* and the *object* and is identified by a URI of the property that models the predicate, and the object which can either be a simple *literal value*

(e.g., string or number) or a URI of another resource. The two possible object types determine two types of RDF triples: *Literal Triples* and *RDF Links*. While literal triples are used to describe properties of a resource, RDF Links describe relationships between two resources. A collection of RDF triples can be also seen as an RDF graph. Since resources on the Web of Data are identified by globally unique URIs it is possible to imagine all Linked Data as one "giant global graph" [5].

There is a tendency that people conceive RDF as a data format which is wrong. It is only a data representation model. In order to publish a collection of RDF triplets (i.e., an RDF graph) on the Web of Data, they must first be serialized into an RDF serialization format. There exist several RDF serialization formats, two of which have been standardized by W3C: RDF/XML and RDFa. So far, the RDF/XML syntax [4] has been the most often used RDF serialization format in spite of the fact that it is difficult for humans to read and write it. One of the main reasons for that is a large number of available XML parsers. RDFa is the RDF serialization format that mixes RDF triples and HTML. It is applicable in contexts where data providers are allowed to modify HTML contents to which they intend to add RDF triples. RDFa embeds RDF triples into HTML as values of a set of attribute level extensions to XHTML. RDFa has risen its popularity mostly on the fact that it enables both human-readable and machine-processable data representations to coexist in the same Web document. Besides RDF/XML and RDFa, other RDF serialization formats include Turtle, N-Triples and RDF/JSON.

4   Semantic Annotation of Environmental Data

RDF is a generic, abstract data representation model for describing Web resources. It does not provide any domain-specific terms (concepts) for describing a group of resources from a certain domain. This task is delegated to *taxonomies*, *thesaurii* and *ontologies* which are usually expressed in languages such as RDFS (the RDF Vocabulary Description Language, also known as RDF Schema) and OWL (the Web Ontology Language). All three *taxonomies*, *thesaurii*, and *ontologies* represent collections of controlled vocabulary terms, each term representing a particular name of one domain concept. They differ in the way the terms are organized as well as the expressivity of relationships between them. A taxonomy organizes terms into a hierarchical structure applying parent-child relationships. A thesaurus is a collection of controlled vocabulary terms that uses associative relationships in addition to the parent-child relationships. An ontology is far more expressive than a taxonomy and thesauri, providing a set of terms, properties (relationships) and property restrictions necessary for the conceptualization of a given domain.

Semantic annotation of linked data refers to the process of linking concepts defined in domain ontologies to RDF nodes that represent the data units to be linked. Semantic annotations, that is annotating by using ontological concepts, represent the conceptualization of information held by the annotated data units.

However, in reality, it is rare that the annotated data units contain exact instance of the annotating ontological concepts unless they belong to the ontology's documentation. It is more realistic that annotated data units model only some aspects of the ontological concepts. Therefore, besides linking the ontological concepts, it would be useful if the annotations also provide information on the relevance (relatedness) of the ontological concepts for the data units they annotate.

The quality of semantic annotation strongly depends on the quality of the domain ontology. Some domains are described by better ontologies than others, but still there is a major shortcoming of standardized domain ontologies. In most cases domain ontologies cover only a specific part of a domain, so that it is very unlikely that data providers can literally reuse an existing ontology. In a more realistic scenario data providers will have to extend some of the existing ontologies by adding new terms that serve their purposes.

4.1 A short "howto" guide for environmental data providers

When choosing the ontology, the environmental data provider should first check the usage of the ontology in the domain. That is, how much the ontology has been used in the domain and whether is has been standardized by an ontology standardization body. Second, they should check if the ontology is actively maintained and updated according to a well-established governance process. Third, does the ontology cover substantial part of the data that the environmental data provider intends to annotate? Finally, they should check if the ontology delivers enough expressivity for their data sets and for the intended application scenarios. After choosing the ontology, if the environmental data providers still need to extend it, here are the aspects that should be considered:

- Identify subsets of concepts and properties of your interest from the ontology;
- Identify missing concepts and properties;
- Define a new namespace that you can control;
- Define new concepts and properties under the defined namespace;
- Document the new concepts with human-friendly labels and comments;
- Relate the new concepts to the existing, related concepts from the ontology;

New concepts can be defined completely from scratch or derived from the existing ones. If derived from the existing concepts, the `rdfs:subClassOf` property should be used to relate new concepts to the parent ones. When relating the new concepts to to some existing and related concepts, an arbitrary number of mappings can be established. The better connected ontology is, the better integrated data that it annotates will be [14]. In case new hierarchical and associative relationships need to be defined, it is recommended to rely on reusing properties from existing well established vocabularies such as SKOS (Simple Knowledge Organization System) rather than defining new ones. For example, the `skos:broader` and `skos:narrower` properties can be used to assert a direct hierarchical link between two concepts, while the `skos:related` property can be used to assert an associative relationship between two concepts. Since

both research on a given environmental domain and the ontologies describing it evolve over time, there is a high probability that at a later point in time the environmental data providers discover that another environmental domain ontology contains the same concepts they have already defined. In that case the `owl:equivalentClass` property can be used to state that the concepts in the two ontologies are equivalent.

Besides choosing and extending the domain ontology, the environmental data providers should also decide on the way ontological concepts are being linked to their data units. This actually means that the environmental data providers should decide on an annotation interface that will provide structures for linking ontological concepts to the data units. We distinguish between two general approaches in designing the annotation interface, each of which having comparative advantages and drawbacks.

The first approach assumes linking concepts from domain ontologies directly to RDF nodes of the data units. This approach enables a very flexible annotation approach and does not require a complex annotation interface. The annotation interface is composed of predefined properties whose domain comprises data units to be annotated and range comprises concepts from domain ontologies. Most commonly used property for this purpose is `dc:subject` which is defined in the Dublin Core (DC) vocabulary. The advantage of this approach is the simplicity of adding/linking annotating ontological concepts to data units. The drawback is limited expressively, which is reflected in the fact that annotating concepts are left without any contextual information such as who added them, when they were added, and how relevant they are.

The second approach assumes the existence of an intermediate annotation concept in the annotation interface, over which domain concepts are linked to the RDF nodes of the data units. This concept is usually named as "Annotation" and provides an arbitrary number of properties that model additional information about the annotating domain concept. Figure 1 shows an example of the annotation interface that conforms to this annotation approach. As it can be seen from the figure, besides the annotating domain concept that determines the annotation, the annotation interface also holds information about the annotation's author, the date and time when the annotation was created, and the relevance (weight) of the annotating domain concept. The last information actually depicts up to what extent the annotating domain concept conceptualizes semantics of the data unit it annotates.

By applying the second annotation approach we have annotations loosely coupled with the data unit, which gives flexibility to add additional information about the annotating domain concept. The main drawback of this approach is that software applications that intend to consume the annotations will need to be aware of the complexity of the annotation interface. Moreover, retrieving the annotations will require the execution of more complex queries, which in case of large collections of interlinked data can cause significant lose of the applications' performance.
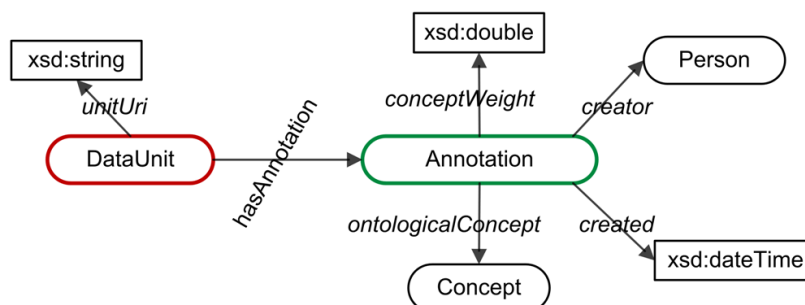
Fig. 1. An example of the annotation interface

## 5   Semantic Linking of Environmental Data

Having environmental data uniquely identified and semantically annotated, the next step in building the semantically unified environmental information space is to create links between related data units. Links should be established both internally, among data units belonging to the same data set (i.e., the data units from the same data provider), and externally between data units belonging to different data sets. Both internal and external links are important in ensuring that the data set is well integrated into the semantically unified environmental information space and further into the global Web of Data. While establishing the internal links is mainly responsibility of the environmental data provider, external links can be established by other environmental data providers and a general Web audience as well. There are two types of external links: *outgoing links* pointing from data units inside the data set to data units from other data sets and *incoming links* pointing in the opposite direction. Both types are equally important in ensuring that each data unit can be discovered and accessed by software applications consuming the data through link traversal.

External links can be established only after the data set is published on the Web. What is important for us is how environmental data providers can contribute to this process. At first glance it might appear that the data providers can only create outgoings links from their data sets. However, when using the RDF data representation model, there is no difference in creating outgoing and incoming links. In both cases, creating links between data units is about generating and publishing RDF triples that describe the links. A good practice would be that the environmental data providers, once they publish their data sets, try to identify related data within other data sets available on the Web and create an initial set of links. Moreover, they should also well document their data sets in order to convince third parties of the value of linking to their data sets.

Considering semantics of a relationship between two linked data units, we distinguish between two types of links: property-based (predefined) and generic links. Property-based links are described by RDF triples in which the subject and object are URIs of the data units to be linked and the predicate is a predefined property that determines semantics of the link. Identifying an appropriate prop-

erty that describes well a desired relationship is a key issue in creating property-based links. In general, it would be better if the environmental data providers reuse properties from existing, well maintained vocabularies (ontologies) than defining new ones. However, if the data providers can not identify suitable properties for intended links, then they can define new ones. This should be done in the same way as defining missing properties in the environmental domain ontology (see Section 4). Generic links model the link's semantics by means of shared semantics of the linked data units. They employ shared, conceptualized semantics of the data units to be liked to determine the link's semantics. Generic links are especially useful in case of automatic link generation as we explain later.

Creating generic links is more complex than creating property-based links. While a property-based link is described by one RDF triple, a generic link is described by a number of RDF triplets that represent an instance of a generic link specification. Figure 2 exhibits an example of the generic link specification. As we can see from the figure, the generic link in this case is specified by the `SemanticLink` class and the following properties: `unitA` and `unitB`, which hold references to the data units linked by the generic semantic link; `linkingConcept` that holds the reference to the ontological concept that conceptualizes shared semantics between the linked data units and determines the semantic relation between them; and the `linkStrength` property whose value determines the strength of the semantic relation, that is, semantic relatedness between the data units.
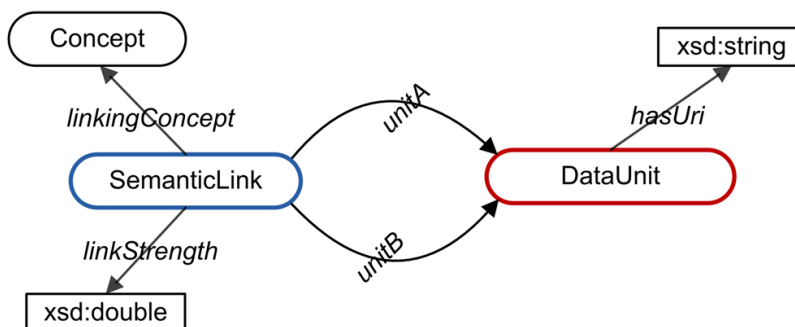
Fig. 2. An example of the generic link specification

Both property-based and generic links can be set manually or automatically. Manual interlinking is more appropriate for small, static data sets, while larger data sets generally require an automated or semi-automated approach. The main idea of the automated approaches is to measure semantic relatedness between the data units and if it is above a given threshold then they generate the links. Generic links appear to be more applicable for the automated link generation than property-based links since they are not constrained to a set of predefined

properties. In principle, possible semantics of the relationships between data units are unbounded so that it is difficult to cover all of them by a predefined set of properties.

6   Conclusions

In this paper we have analyzed the requirements that environmental data providers should satisfy when publish their data, so that the data becomes seamlessly integrated into LOD cloud and the Semantic Web. The requirements consider environmental data identification, representation, annotation and linking. Publishing environmental data according to the proposed requirements will lead towards creation of the semantically unified environmental information space. Search engines and intelligent software agents will take advantage of such unified information space to enhance discovery of environmental data and to assist humans in improving their research and understanding of the environment we live in.

Acknowledgements

References

1. I. N. Athanasiadis and S. Janssen. Semantic mediation for environmental model components integration. *Information Technologies in Environmental Engineering*, 1(3-11), 2008.
2. I. N. Athanasiadis, A. E. Rizzoli, M. Donatelli, and L. Carlini. Enriching environmental software model interfaces through ontology-based tools. *International Journal of Applied Systemic Studies*, 2011 (In Press).
3. C. Becker and C. Bizer. Exploring the geospatial semantic web with dbpedia mobile. *Journal of Web Semantics*, 7(4):278–286, 2009.
4. D. Beckett. RDF/XML Syntax Specification (Revised) - W3C Recommendation. http://www.w3.org/TR/rdf-syntax-grammar/, 2004.
5. T. Berners-Lee. A Framework for Web Science. *Fondation and Trends in Web Science*, 1(1):1–130, 2006.
6. C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The story so far. *Int. Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.
7. J. Hendler. Agents and the Semantic Web. *IEEE Intelligent Systems*, 16(2):30–37, 2001.
8. I. Jacobs and N. Walsh. Architecture of the World Wide Web. http://www.w3.org/TR/webarch/, 2004.
9. S. Janssen, I. N. Athanasiadis, I. Bezlepkina, R. Knapen, H. Li, I. P. Dominguez, and A. E. Rizzoli. Linking models for assessing agricultural land use change. *Computers and Electronics in Agriculture*, 2010.

10. S. Janssen, F. Ewert, H. Li, I. N. Athanasiadis, J. Wien, O. Thérond, M. Knapen, I. Bezlepkina, J. Alkan-Olsson, A. Rizzoli, H. Belhouchette, and M. Svensson. Defining assessment projects and scenarios for policy support: use of ontology in integrated assessment and modelling. *Environmental Modelling and Software*, 24:1491 – 1500, 2009.
11. W. Lloyd, O. David, J. Ascough, K. Rojas, J. Carlson, G. Leavesley, P. Krause, T. Green, and L. Ahuja. Environmental modeling framework invasiveness: Analysis and implications. In *International Congress on Environmental Modelling and Software (iEMSs 2010)*, pages 1073–1080, 2010.
12. J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa. An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, 2(3):279 – 296, 2007.
13. R. Moats. Rfc 2141: Urn syntax. http://tools.ietf.org/html/rfc2141, 1997.
14. S. Nešić, F. Crestani, D. Gašević, and M. Jazayeri. Search and Navigation in Semantically Integrated Document Collections. In *Proceedings of the 4th International Conference on Advances in Semantic Processing, SEMAPRO*, pages 55 – 60, Florence, Italy, 2010.
15. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.
16. A. E. Rizzoli, M. Donatelli, I. N. Athanasiadis, F. Villa, and D. Huber. Semantic links in integrated modelling frameworks. *Mathematics and Computers in Simulation*, 78(2-3):412 – 423, 2008.
17. F. Villa. Integrating modelling architecture: a declarative framework for multi-paradigm, multi-scale ecological modeling. *Ecological Modelling*, 137:23–42, 2001.
18. F. Villa, M. Ceroni, K. Bagstad, G. Johnson, and S. Krivov. Aries (artificial intelligence for ecosystem ser- vices): a new tool for ecosystem services assessment, planning, and valuation. In *11th International BIOECON Conference*, 2009.
19. F. Villa, M. Ceroni, and S. Krivov. Intelligent databases assist transparent and sound economic valuation of ecosystem services. *Environmental Management*, 39:887–899, 2007.