Semantic Bookmarking and Search in the Earth Observation Domain

Francesca Fallucchi[1], Maria Teresa Pazienza[1],
Noemi Scarpato[1], Armando Stellato[1],
Luigi Fusco[2], and Veronica Guidetti[2]

[1] DISP, University of Tor Vergata,
Via del Politecnico 1, Rome, Italy
{fallucchi,pazienza,scarpato,stellato}@info.uniroma2.it
http://ai-nlp.info.uniroma2.it
[2] ESA/ESRIN, Frascati, Italy
{luigi.fusco,veronica.guidetti}@esa.int

Abstract. This paper describes the experience of introducing a service for semantic bookmarking and search in the Earth Observation (EO) domain. To perform the work reported such a service has been integrated and customized in the framework of the DILIGENT project which concluded in 2007. The service proposed is a web browser ontology based extension: at the application layer the interaction between DILIGENT end-users and this service takes place from within the DILIGENT web portal. The service described allows end users to capture and annotate on-demand the information stored in the DILIGENT infrastructure and to organize it according to the user's representation of the specific ontological domain. The effort spent meets today end users' requirements in the EO application domain, applying an innovative solution to make specific actors keep track of the information of interest in a personalized and easy way. End users are able to store, access and play only with the information of interest and under their own (ontological) perspective.

1 Introduction

Nowadays, images of our planet from orbit are acquired continuously; they have become powerful scientific tools to enable better understanding and improved management of the earth and its environment. Space taken Earth Observation (EO) images show the world through a wide-enough frame so that complete large-scale phenomena can be observed with great accuracy. EO provides objective coverage across both space and time. The same space-based sensor gathers data from sites across the world, including places too remote or otherwise inaccessible for ground-based data acquisition. In the long term, this monitoring of the earth's environment will enable a reliable assessment of the global impact of human activity and the likely future extent of climate change. Due to the flexibility (and complexity) of the instruments, it has to be noted that one single EO instrument supports many different domains (e.g. ocean and land application) and, at the same time, one application domain needs to access many different sources of information from space (EO) and from ground measurements.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part III, LNAI 5179, pp. 260 – 268, 2008.

Another important aspect to note is that different space and ground instruments are operated under the responsibility of different entities (e.g. space agencies, research institutions…), so that the interoperability of the different available data systems to generate data products and information to users is not at all a trivial task. Definitely, the semantic web technology can play a very important role to easy the various intrinsic interoperability problems for data discovery, access and use. As immediate example of the help that semantic technology can help, we refer to the handling of time and geographic location associated with each image dataset, which can be expressed in many different ways, units, relative to other events etc.

Furthermore, the user services which ensure data and information access, like those offered and shown through the ESA's EO Portal[1], have to deal with continuous streams of data coming from satellites, which need to be collected, organized and presented to their actual and potential consumers. Managing this data is not a trivial task, not only for the huge quantity of available material, but also considering the several, diverse scenarios in which its content can be used as well as the different categorization schemas which can be thought over it. So far, providers of geographical information have archived their knowledge in huge databases and found complex projections of their raw data for populating web sites and portals with rich and navigable descriptions of their content. Specific consumers' exigencies are instead been dealt, case by case, with the knowledge and expertise possessed by educated and trained personnel.

Modern web technologies, pushed by Web 2.0 and its leading objective of a read/write Web, and strong knowledge representation standards offered by the Semantic Web [2], offer now viable alternatives which could be used to fill the gap between data archiving and data publication, by allowing for the latter to being driven through proper, multi-modal representations of the raw data as offered by ontology based approaches. RDF/OWL marked-up versions of the acquired data could not only provide a layer which can be easily ported to the web, but may offer diverse perspectives on the same raw data which can thus be easily aggregated, selected and composed according to different needs and exigencies.

2 Context and Scenario

ESA/ESRIN reached today several years of world-wide experience in cooperating on ICT topics concerning the exploitation and integration of digital libraries and Grid technologies, putting the basement to exploit such technologies in earth science and EO domains. DILIGENT[2] is one of the European projects in which ESA/ESRIN participated during the past three years with the role of end-user and services/data provider, exposing requirements from the EO community to gain from innovative ICT based solutions.

The test-bed scenario lead in the context of DILIGENT based on concepts like virtual organization, digital library, user workspace, collection, compound service, report with the aim of creating on-demand ad hoc digital libraries and then aggregating

[1] http://www.eoportal.org/
[2] http://www.diligentproject.org/

pertained information into complex report documents (e.g. reports on the environment status), browsing, searching and managing private workspaces from within *virtual research environments*. During the project lifetime, together with the furthering of the activities and the development of the DILIGENT platform through the definition of its middleware *gCube*[3] and a pan-European Grid infrastructure, the need to exploit knowledge representation techniques, annotation and ontology management raised up in the context of the EO scenario.

To this end ESA/ESRIN required the support of University of Rome, Tor Vergata, which shown a certain experience in the development of ontology based service for web pages annotation. In the scope of DILIGENT such a joint effort moved towards two directions: a) to define a suitable ontology for demonstration purposes in the project and b) to integrate the ontology based service in gCube. The former point involved the definition of an ontology based on the ESA's EO Portal structure together with the evaluation of ontology publicly available in the EO domain, as the SWEET ontologies developed by NASA [12], while the latter saw the proposal and implementation of a technical solution as described in the next sections.

The main motivation which lead the project, and the ESA participation in particular, in this direction resides in the lack of a user-centric approach in the majority of web portals and access points for EO information and services: currently, people enter the EO portal and browse its content like in any standard web site; the portal offers a few search functionalities and its standard html pages are evidently produced automatically from templates + content regularly fed from a database. The EO user portal, which is being developed in DILIGENT, is based on slightly more recent technologies, essentially dynamic html, and features interaction modalities based on "content objects": they are moveable interactive knowledge units, generated upon searches submitted by the user, which can be inspected, aggregated and moved into sort of bookmark folders which are saved and accessed on a per-user basis. In both cases however, users could beneficiate of the underlying knowledge organization which completely drives the structure of both portals.

Unfortunately, in our present context, we were limited to research on a lightweight solution providing interesting features for potential users of the EO portal, without requiring heavy modification or introduction of new components inside the business logic of the systems behind the portal: we thus had to concentrate on some sort of customizable desktop solution, able to provide fast content collect&retrieve functionalities for browsed data, as well as maintaining a strong connection with the source where it has been extracted. This rich client would have dedicated functionalities thought (or customized) for the eoPortal, though not relying on any kind of preferential access to its content, but standard http interaction.

3 Approach

Given the above constraint, which required no substantial change in the existing system, we decided to adopt a completely client-side navigation mechanism: a browser-on-semantic-steroids, which, beneficiating of the expressive power of Semantic Web

[3] http://www.gcube-system.org/

W3C formalisms for providing both ontology editing and annotation/bookmarking functionalities, could at least support users in the process of collecting and organizing information observed through traditional navigation in the portal, implicitly maintaining references to its resources. Ontologies assume thus a two-fold role in this approach, on the one side, the domain ontology established by the user, on the other hand, a mere ontological translation of the original metadata defined inside the eoPortal. With this approach, and by developing the platform described above, users could depict their personal perspective on the interested data, implicitly defining filters, projections and data transformations, through the act of developing an own representation of their specific domain of interest. This representation would remain deeply coupled with the underlying standard eoPortal data, for search purposes.

Following these premises, we defined the following two working steps:

- the first requiring no intervention on the portal, with a client application just being "aware" of portal content, structure and navigation modalities, and allowing for collecting and organizing its content according to user needs.
- The second step involves minor operations on the portal side, by allowing the explicit exposure of metadata inside the same pages which are browsed through our enhanced client, which can then be easily harvested and reorganized into the client platform

Fulfilling the above steps required, first of all, to design the rich client platform, or to delineate its characteristics and then identify a proper candidate for providing the new eoPortal service.

3.1 The Client Platform

A recent work from the University of Tor Vergata, the Semantic Bookmarking platform Semantic Turkey [8], was elected by ESA/ESRIN as a potential application for achieving their objectives. Its main advantages with respect to other possible candidates were: the fact that it is based on a real Web Browser (Mozilla Firefox), thus guaranteeing robustness with respect to any sort of document which can be found on web pages, its native datamodel, based on the OWL language, and its flexible extension mechanism (based on the standard OSGi), which left the door open for customization to ESA needs.

Semantic Turkey (ST), in its original version, is a "Semantic Bookmarking" platform: an hybrid between a Web Browser, an Annotation tool and and Ontology Editor (see fig. 1). The expression Semantic Bookmarking was coined to indicate the process of annotating information from (web) documents, to acquire new knowledge and represent it through representation models. Its basic functionalities allow for:

1. capturing information from web pages – both by considering the page as a whole, as well as by selecting portions of their text – and annotating it with respect to a personal ontology.
2. editing the above ontology for classifying the annotated information and for better characterizing its interests according to its descriptive properties (attributes and relations).
3. navigating the structured information as an underlying semantic net which, populated with the many relationships which bind the annotated objects between them, eases the process of retrieving the knowledge which was buried by the past of time.

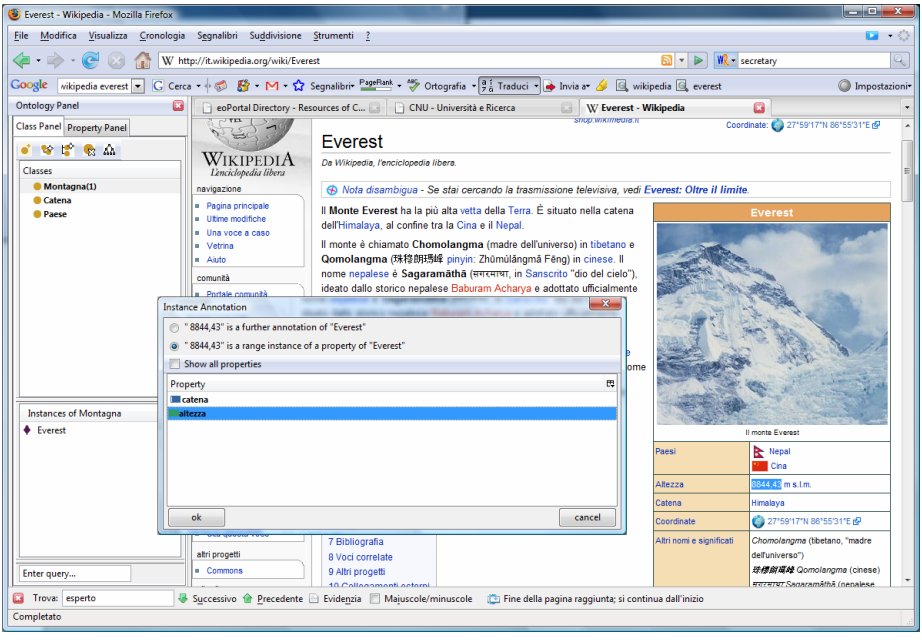

Fig. 1. Semantically bookmarking a reference ontology with objects from web pages

Its architectural and functional design make Semantic Turkey differentiate from similar, existing annotation tools, as it offers a lightweight structure, which completely exploits the infrastructure of the hosting web browser (with respect to, for example, the complex completely-web based interface of Piggy-Bank [9]) and which grants the user a good control over its personal domain representation (while traditional semantic annotation tools like Magpie [6] and Melita [4], are only able to import and adopt ontologies which have been defined elsewhere).

3.2 Preserving the Source Metadata

The data used to produce pages of the eoPortal is organized after a general-purpose model for resources management and documentation, reported and documented in [13]. This model has been implemented inside a Database Service (details in [11]) which hosts both the pure metadata entries as well their processed information which is directly accessed for populating the presentation pages of the eoPortal.

In Semantic Turkey there is a neat separation between the explicit knowledge managed by the user (*user layer*) and the one which guides system's behavior (*application layer*). The role of this latter layer is to keep track of all the information which is required by the application to perform its functionalities and is typically hidden form the user; in the original version of ST it includes the *Annotation Ontology*, a small ontology containing a set of concepts used to keep track of user annotations from the web (annotated web pages, selected textual entries, timestamps etc…). In order to provide an homogeneous data layer for our framework which is based on technologies and languages of the RDF family, we realized an almost straight 1-1 porting of the

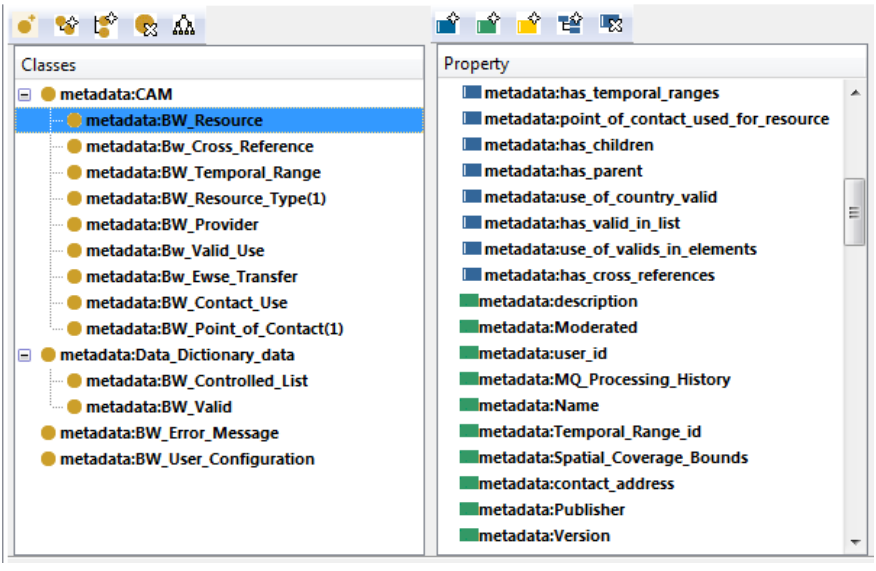

Fig. 2. Partial view of the ESA metadata in their straight ontological translation

eoPortal metadata model to an ontological representation (see fig. 2) and added it to the application layer of Semantic Turkey. This preserves consistent reference to the eoPortal native data model while guaranteeing interoperability with the specific ontologies that different users may adopt for their personal semantic bookmarks.

3.3 Automatic Data Harvesting: Exploiting RDFa

The second step of our effort on introducing Semantic Web technologies in the EO domain went in the direction of making data explicit (whenever possible and if available) inside the same traditional web pages used for presentation: we have thus considered the new possibilities offered by RDFa and Microformats .

Microformats (http://microformats.org/) emerged from the work of blog aggregator Technorati's developer community (http://www.technorati.com/), following the principles of the Global Multimedia Protocols Group (http://gmpg.org/): they are a proposed format for making recognizable data items (such as events, contact details or geographical locations) capable of automated processing by software, as well as directly readable by end-users. Microformats consist in specific HTML extensions intended for carrying commonly published semantics, such has contact information, events, reviews, episodic content, which go beyond the range of their hosting language. RDFa [1, 10] is the natural answer from W3C, associated to Semantic Web RDF-based models and technologies. The key aspect in both technologies is the same: embedding data inside web pages, which is exactly what we needed to achieve. The main difference (syntactic sugar apart) between the two resides in the vocabulary generation. While there exists one specific microformat for each of the addressed domains (contacts, events etc…), the only existing RDFa vocabulary is based on the

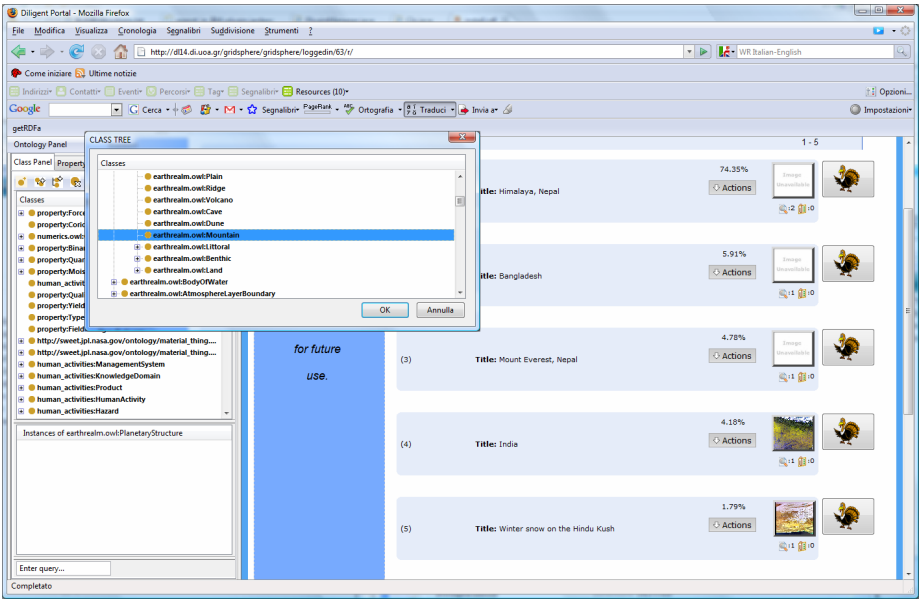

Fig. 3. recognizing, extracting and bookmarking RDFa data in the DILIGENT portal

RDF family of web languages (RDF, RDFS, OWL), thus extending existing HTML tags with attributes like: instanceof, property, etc.. and relying on specific content (the values of these attributes) driven by referenced ontologies.

We have thus developed an extension for Semantic Turkey specifically designed for recognizing information coded inside web pages according to these languages, and use their content both for populating the user ontology or, again, for submitting it to the semantic repository. Due to its W3C standardization and its flexibility with respect to microformats, we chose RDFa as the main language whose content can be recognized and extracted from web pages, and mapped to the user ontologies. For allowing a wider compatibility with available data on the web, we identified those microformats which could fall in the domain of Earth Observation, like: hCalendar (http://microformats.org/wiki/hcalendar), for describing time events, and geo (http://microformats.org/wiki/geo), for marking up WGS84 geographic coordinates (latitude and longitude) and enabled the conversion of their content through dedicated GRDDL (Gleaning Resource Descriptions from Dialects of Languages [5]) gleaners, to match the specification of standard W3C ontologies (hCalendar is by default translated to iCal, http://www.w3.org/2002/12/cal/ical, while geo is converted to WGS84 ontology [3]).

Fig. 3 shows a typical user session with RDFa data: the page in the browser comes from the DILIGENT portal, and contains data coded in RDFa after an OWL translation of the DILIGENT Metadata model. Semantic Turkey has recognized these data and added buttons (they are on the right, with the turkey symbol on top) in correspondence of their position in the web page. The user can thus press the buttons and get all the related RDFa (it generally involves multiple RDF triples describing a single

object) collected in one click: he can then decide if the data will be imported as they are (i.e. referring the ontology of the RDFa data source), or if they will be projected upon the domain ontology (like the SWEET ontologies showed in figure) adopted by him. In the latter case, each object retrieved from the RDFa description can be projected towards a concept of the user domain ontology. 1-to-1 or directed 1-to-many mappings between concepts and properties of the ESA Metadata ontology and those of the domain ontology adopted by the user can be defined a-priori or they can be declared "inline" for each new harvested object. Property reification may also be used to map properties of one ontology towards classes and roles of the other one. This mapping capability has proved to be sufficient in most of the case, partially facilitated by the lack of specification on data format present in many currently available ontologies (such as the already cited SWEET ontologies). Actually, in our case study, predefined mappings proved to be helpful mostly for submitting info to the portal (though this aspect has not been investigated extensively), while their usefulness were limited to very few cases (for example, the already cited calendar or geospatial ontologies and microformats) when projecting harvested data from the web to the user ontology. The main reason for that resides in the ESA Metadata ontology, which contains few top-level concepts providing a perspective focused over mere data organization, with no evident relation to domain information, so that it was pretty easy to define in advance data projections from the rich user domain ontologies towards the general concepts of the ESA model, while relying on on-the-fly mappings for doing the contrary.

4 Conclusions and Future Work

This paper reports on experimenting with technologies coming from both worlds of Web 2.0 and the Semantic Web in the domain of earth science and EO specifically

Our research in this field will continue inside the new D4Science project (http://www.d4science.org/) - which started at the beginning of 2008 as a natural follow-up of DILIGENT - and will focus on improving ontology based content organization, and search and retrieval functionalities even on the service side.

We will probably extend current mapping capabilities of the framework, mainly a) to extend the range of achievable mappings, also including raw datatype transformations and more complex mapping patterns and b) to facilitate the user in harvesting several objects and in relating them to the adopted domain ontology, by exploiting results from pattern based ontology design [7] and applying them to our data projection facility. These additions (especially for the second purpose), will require methodological analysis of user activity on a range of realistic use-cases, considering different domain ontologies as a test basis, and keeping track of desired mappings which cannot easily be projected according to available models. User-friendliness, on the other side, has to be taken into account (though hardly measurable if not through user questionnaires), to avoid approaches proving to be potentially "more complete", but of unrealistic applicability in the hands of the average user.

Moreover a deeper attention will be paid to the exploitation of the ESA's EO Portal that aims to open the door to the world of EO resources, in its role of single access point to EO information and services provider.

Acknowledgements

The joint effort described in this paper has been partially funded by DILIGENT (Call Identifier: FP6-2003-IST-2 Project number: 004260).

References

1. Adida, B., Birbeck, M.: RDFa Primer. Retrieved from W3C (October 26, 2007), http://www.w3.org/TR/xhtml-rdfa-primer/
2. Berners-Lee, T., Hendler, J.A., Lassila, O.: The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American 279(5), 34–43 (2001)
3. Brickley, D. (n.d.).: Basic RDF Geo Vocabulary (Retrieved December 10, 2007), http://www.w3.org/2003/01/geo/
4. Ciravegna, F., Dingli, A., Petrelli, D., Wilks, Y.: User-system cooperation in document annotation based on information extraction. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, Springer, Heidelberg (2002)
5. Connolly, D.: Gleaning Resource Descriptions from Dialects of Languages (GRDDL). Tratto da World Wide Web Consortium (September 11, 2007), http://www.w3.org/TR/grddl/
6. Dzbor, M., Motta, E., Domingue, J.B.: Opening Up Magpie via Semantic Services. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 635–649. Springer, Heidelberg (2004)
7. Gangemi, A.: Ontology Design Patterns for Semantic Web Content. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M. (eds.) Proceedings of the Fourth International Semantic Web Conference, Galway, Ireland. Springer, Heidelberg (2005)
8. Griesi, D., Pazienza, M.T., Stellato, A.: Semantic Turkey - a Semantic Bookmarking tool (System Description). In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, Springer, Heidelberg (2007)
9. Huynh, D., Mazzocchi, S., Karger, D.: Piggy Bank: Experience the Semantic Web Inside Your Web Browser. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 413–430. Springer, Heidelberg (2005)
10. Lassila, O., Hendler, J.: Embracing Web 3.0. IEEE Internet Computing (May-June 2007)
11. Pedersen, G.S.: Database Design Document. Centre for Earth Observation (CEO). CEO programme Office (TBD copies) (1999)
12. Raskin, R.: Semantic web for earth and environmental terminology (2005), http://sweet.jpl.nasa.gov/index.html
13. SSSA, Recommendations on Metadata: Describing the data, services and information you have available! Ispra (VA), Italy: Strategy and Systems for Space Applications Unit (SSSA), Space Applications Institute (SAI) / Joint Research Centre, European Commission (2000)