Journal of Web Engineering, Vol. 2, No. 4 (2004) 295–304
© Rinton Press

STRUCTURING INFORMATION ON THE WEB FROM BELOW: THE CASE OF EDUCATIONAL ORGANIZATIONS IN CHILE[a]

Ernesto Krsulovic-Morales
*Department of Computer Science, Universidad de Chile, Blanco Encalada 2120
Santiago, Chile
ekrsulov@dcc.uchile.cl*

Claudio Gutierrez
*Department of Computer Science, Universidad de Chile, Blanco Encalada 2120
Santiago, Chile
cgutierr@dcc.uchile.cl*

Received March, 2004
Revised October 29, 2004

This paper reports the framework and the experience of structuring and integrating information of educational organizations in Chile, using metadata along with Semantic Web ideas. We present an implementation for Computer Science departments and a more general framework for educational organizations.

Keywords: Semantic Integration, RDF, Semantic Web, Yearbook, Education

Communicated by: R Baeza-Yates

1 Introduction

One of the most attractive aspects of the Semantic Web (SW) is the promise to be able to make inferences over the information on the Web resulting from the integration of applications [1]. This promise will not be realizable until the basic architecture proposed by the W3C is implemented and we have a critical mass available to allow its utilization [2]. While waiting for these days to come, we can take advantage of the most basic and stable part of this architecture, namely the metadata layer (standardized by the RDF model), to create applications which require more structured information, but with distributed providers. There are several tools that "structure" information, stores it, and allow users to query that information. The three main paradigmatic models are classical databases, directories, search engines; but there are several others with particular features, e.g. yearbooks (constraint of time added), intranet search engines (constraint on the set of URIs visited), etc. Figure 1 is a comparison among different solutions. When the requirements are distributed editing, different schemas, and lack of information about the final use of the information, an ontology-driven model is a very attractive option. We present in this paper a case of integration based on this model for the educational area.

[a] Supported by Millennium Nucleus "Center for Web Research" (P01-029-F), Mideplan, Chile.

Features	**Database**	**Directory**	**Search Engine**	**Ont-driven**
data model	schema fixed	schema distr. edition	none	ontology distr. edition
data location	centralized	centr/distr	centralized	distributed
access level	views	views	views	data source
data load	manual/ autom. schema asssisted	manual schema assisted	crawler	manual ontol. assisted
suscribtion	requested	submitted	web opt-out	web opt-in
search	query SQL-like	browsing patt. match.	patt. match.	RDQL patt. match.
model ext.	sufficient	bad	good	very good
trust level	very good	good	bad	bad

Fig. 1. Comparison of different forms of structuring information on the Web

The educational area has been historically an ideal area for concept-proof of new technologies, due to the low risk level, the openness of the people to adopt new ideas, and the need of solutions that support their activities. Curiously, widely deployed metadata initiatives are scarce in the educational area. Currently, metadata initiatives focus mainly on the integration of educational contents. There are good examples at a national level in the area of e-learning, like TheGetaway (USA), Plash (Canada), Edutella and Universal (Europe), which aim at the use of standard metadata for learning objects [3] to be retrieved in portals or over P2P networks. There seems to be no successful initiatives of information integration based on metadata and ontologies for educational organizations.

In Chile, in order to integrate educational resources on the Web,[b] almost the only choice available is to use a search engine, e.g. *TodoCl.cl*, with all the limitations involved. Other approaches currently used are the building of communities around portals, the so called *global gateways*, which concentrate the information under a common roof, implying centralization of the data model as well as the information. This approach has several drawbacks compared to the use of metadata [4]. Two examples in Chile are *EducarChile.cl*, focused on school level, and *Universia.cl*, focused on the university level. Almost all of them are built using LDAP or relational databases. Our present work presents an alternative distributed approach using metadata.

In this paper we present a project of information integration on the educational area in Chile: a yearbook for the Chilean CS Departments, and its extension for educational organization at large. The driving goals of this work were: (1) To help populate the Web with metadata by structuring and integrating information of organizations at a small scale. This is a natural complement to big scale projects to build the Semantic Web infrastructure; (2) To give a distributed integration tool to the Computer Science departments in Chile; (3) To provide a framework to extend this work to the educational area in Chile.

The paper is organized as follows: We present preliminary material about RDF and yearbooks in Section 2. Then we present the case of the Chilean Compute Science Departments in Section 3. Section 4 is devoted to educational organization in general, and conclusions are

[b] The target market for educational information in Chile is not small for local standards. In Chile today there are 11.066 schools and 236 universities and colleges distributed in 13 regions. Most of them have a Web site.

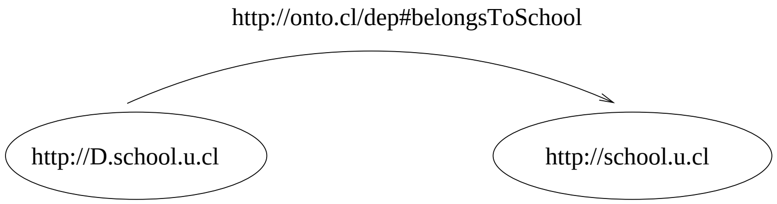


Fig. 2. Example of an RDF graph

presented in Section 5.

2 Preliminaries

RDF. *RDF (Resource Description Framework)* is the metadata model and framework proposed by the W3C, and is rapidly gaining popularity as a de facto standard [5]. Using RDF instead of plain XML (not to mention HTML) offers several advantages: (1) a better model from a semantic point of view; (2) full semantic extensibility; (3) facilities to handle distributed data.

A relational or XML database has a very static schema and it is very difficult to make any significant changes without impacting on existing code. Compared to XML, the RDF model is dynamic and it is easy to add new information to an existing description. Moreover, when keeping metadata with RDF it is possible to scale solutions by integration of previously disjoint organizations.

RDF offers a better semantic model than existing alternatives due to the simplicity of its model (a labeled graph where order is not relevant) and its flexibility to describe data (a simple subject-predicate-object model) [6]. In RDF, resources are identified by a Uniform Resource Identifier (URI), an identification schema which generalizes URLs. In an RDF statement predicates as well as subjects are URIs and the object is a URI or a literal. So for example, it is possible to register the statement "Department D belongs to School S" with the following triple:

```
<http://onto.cl/dep#belongsToSchool>
<http://D.School.u.cl/>
<http://School.u.cl/>
```

where the first is the predicate's URI indicating to what School belongs the given department, the second identifies the department, and the third URI identifies the School. The graph corresponding to this statement is shown in Figure 2.

To describe information on the Web with RDF, we need a common vocabulary, i.e. an *ontology*. One such example is Dublin Core, a minimal vocabulary to describe title, author, copyrights, etc. of a document [7]. For our purposes we need more specific ontologies. RDF allows the specification of such ontologies through RDF Schema (RDFS), an extension of RDF which incorporates functionalities to describe classes, sub-classes, hierarchies, relations and properties of classes [8]. For example, in our prototype yearbook which models university Departments, we need to model people who belong to a Department and their contact addresses, obtaining a hierarchy of classes, attributes of the class *persona* and the relations among classes as shown in Figure 3.

Handling distributed data with RDF is simple because, as described above, an RDF speci-

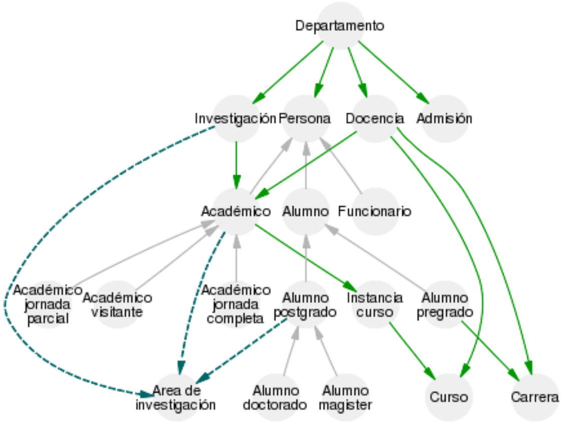


Fig. 3. Main classes of the ontology for Depmark. Arrows pointing up indicate subclass, and pointing down indicate attributes. For example, *Alumno pregrado* (undergraduate student) is a subclass of *Alumno* and has attribute *Career*

fication is a collection of atomic sentences over a shared common vocabulary built by different parties. Storing and querying atomic sentences over diverse vocabularies can be a problem when the volume of data is huge. Currently this issue is a very active area of research (cf. [21]) and resembles the problem of storing efficiently relational data in the early seventies or XML data at the end of the nineties.

Yearbooks. A yearbook is a periodic document issuing information (reports, statistics, etc.) about a particular subject, and is becoming a common practice to publish them in digital format on the Web. Examples are annual reports published by companies refering to departments, projects, financial information. From a semantic point of view, a classic yearbook is directed to a restricted set of users and usually the process of consulting it is mainly a human-oriented task. Building yearbooks is a typical task in many organizations which, although straightforward at first sight, becomes complicated when the organization has strong hierarchies or several horizontal components.

The classic approach to building these information centers is to delegate the task to one of the associated members, which becomes in charge of designing or updating the schema of the data, discussing it with other members, collecting the information among the different components of the organization, and populating the database.

To end this introductory discussion, let us make some remarks about the difference between a yearbook and a catalog in our context and some tips learned about temporality of metadata. A catalog is a list of items, usually ordered, whose goal is describing data in a succint way, e.g. exhibition catalogs, library catalogs, shopping catalogs. The main implicit characteristic of a classical catalog is the hidden assumption about the invariability (atemporality) of the data it describes. (This is more obvious in the shopping catalogs, where price, season, etc. are key parameters). On the Web, catalogs turned exactly into its opposite: although they still do not have a temporal parameter, its information –in the form of links to web sites– is highly variable. In fact, any attempt to describe data on the Web faces the problem of (extremely and unpredictable) variability of data. Incorporating the issue of temporality is a

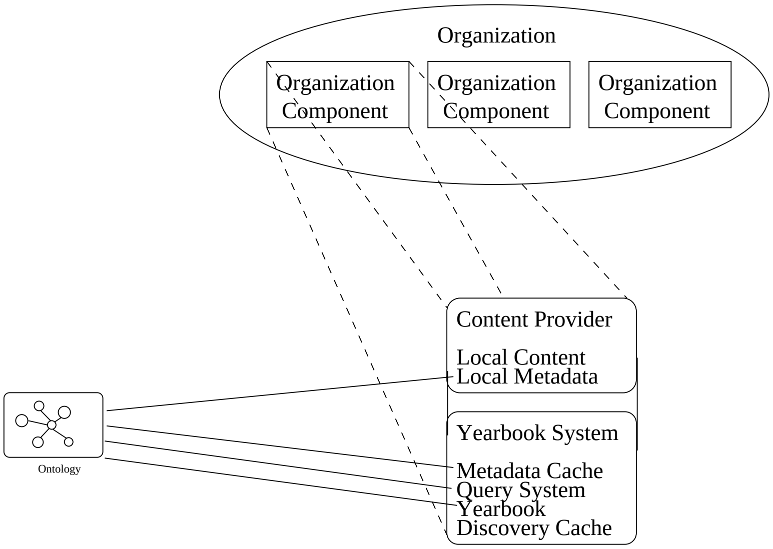


Fig. 4. Yearbook system and sites of the organization components

challenge we are not facing in this work.

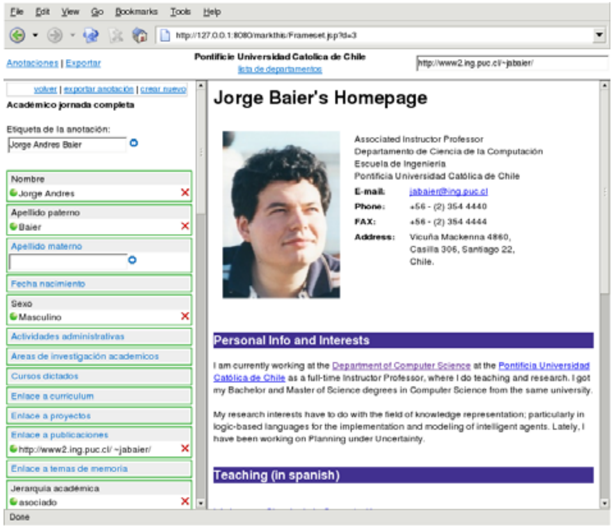
3 The Case of Computer Science Departments

Our work focuses on organizations with small number of components: We are targeting educational organizations in Chile (Schools, Universities, Institutes, etc.) We chose the educational area because, besides the intrinsic value for users, it has many advantages for our research: a known environment, no direct commercial interest involved, people are open to adopt and test new technologies, and permanence over time.

We assume two important weaknesses of today's Semantic Web (SW): (1) the fact that several of the language specifications involved in the SW are not yet stabilized, and (2) the lack of accessible tools to facilitate the manipulation of metadata, especially regarding markup of resources by non-expert users. It seems that we are still far from a common platform between applications for the SW, like "*libwww*" and the browser "*mosaic*" in the beginnings of the Web. There are some interesting attempts, for example Jena [9], a toolkit to build applications with RDF in Java, and Haystack [10] to help markup and retrieval of resources.

The system follows a standard architecture for several SW applications: An *ontology*, a *site markup tool*, a *collect of metadata system*, and a *query system*. Figure 4 sketchs the structure, whose details are discussed in [11]. The whole system must be lightweight and portable to allow its deployment even in components which could not have the infrastructure necessary to host big systems. To cope with the goals proposed, tools should be web-based and oriented to users with no experience in knowledge representation techniques.

Although Academia has been used before as a test bed for markup projects, and there are several university ontologies available (see e.g. [20]), the differences with the Chilean university organization made it advisable to create yet another one. In fact, the solution we found was to build a local ontology (a view is shown in Figure 3), and then specify, at the right level, the conceptual translations to other university ontologies. Our ontology can be


File
Edit
View
Go
Bookmarks
Tools
Help
http://127.0.0.1:8080/markthis/Frameset.jsp?id=3
Anotaciones | Exportar
volver | exportar anotación | crear nuevo
Done

Pontificie Universidad Catolica de Chile
lista de departamentos

http://www2.ing.puc.cl/~jabaier/

Académico jornada completa

Etiqueta de la anotación: Jorge Andres Baier

Nombre: Jorge Andres

Apellido paterno: Baier

Apellido materno:

Fecha nacimiento

Sexo: Masculino

Actividades administrativas

Areas de investigación academicos

Cursos dictados

Enlace a curriculum

Enlace a proyectos

Enlace a publicaciones: http://www2.ing.puc.cl/ ~jabaier/

Enlace a temas de memoria

Jerarquia académica: asociado

Jorge Baier's Homepage

Associated Instructor Professor
Departamento de Ciencia de la Computación
Escuela de Ingenieria
Pontificia Universidad Católica de Chile

E-mail:	jabaier@ing.puc.cl
Phone:	+56 - (2) 354 4440
FAX:	+56 - (2) 354 4444
Address:	Vicuña Mackenna 4860, Casilla 306, Santiago 22, Chile.

Personal Info and Interests

I am currently working at the Department of Computer Science at the Pontificia Universidad Católica de Chile as a full-time Instructor Professor, where I do teaching and research. I got my Bachelor and Master of Science degrees in Computer Science from the same university.

My research interests have to do with the field of knowledge representation; particularly in logic-based languages for the implementation and modeling of intelligent agents. Lately, I have been working on Planning under Uncertainty.

Teaching (in spanish)

Fig. 5. Depmark Markup Tool

accessed at [12].

As test bed we are targeting Computer Science departments at Chilean universities (See project at `http://purl.org/net/depmark`). This area has several advantages for testing such a tool: besides the intrinsic value for users, a known environment, no direct commercial interest involved, people open to adopt and test new technologies, and permanence over time. Moreover, all of them have Web sites, constantly updated, and with rich and diverse information in content and format. The central problem we faced implementing the above architecture was the non-existence of a markup tool. Without such a tool it would be very difficult for people that generate contents (professors, researchers, students, directors, employees, etc.) to generate the metadata about current information on Web pages.

Depmark has two components driven by the ontology: a Web-based markup tool and a yearbook system (see Figure 4). The markup tool was implemented in JAVA. As API to manage RDF models we used JENA. The storage is implemented on the RDB interface model provided by JENA using MCKOI as relational database (a database implemented in JAVA that can be embedded in the application). The schema is RDFS serialized in XML. The markup interface is implemented in JSP.

4 The Case of Educational Organizations

Some of the most important initiatives in the development of educational metadata standards are refered to in [13], some of which are: Dublin Core Education Working Group, IMS Global Learning Consortium, Canadian Core Learning Resource Metadata Protocol, IEEE LTSC-LOM. A project at school level which is a good example of systems based in metadata is the European Treasury Browser [14], which is described as follows: "The goal of the ETB project is to build an infrastructure of an *online educational resource network* for European schools, by integrating existing national repositories, promoting new publications and providing quality levels and trustable structures."

The majority of the components of the Chilean educational network has a Web site. Hence it is possible to make their information inteoperable using standard search engines, for example, a documentary database. Nevertheless, this type of integration would give only the structure provided by the HTML level, which in Chilean sites essentially points to structure the visualization of information. Another approach is given by communities created around portals. There are at least two important at national level: EducarChile.cl, which focuses in the school level, and Universia.cl, which focuses in the university level. A discussion about the advantages and disadvantages of using metadata in these communities can be found in [4].

Our target is the "pre-university" level, that is, those resources for people that are in the last year of high school planing to enter to a university or technical institute. The project describes the market characteristics of educational institutions at that level, such as admission requirements, geographic location, majors offered, etc. The framework extends the Depmark project to a general framework by creating an extension of *profile* [15] of one of the existing standards. Additionally, it develops a tool that facilitates the composition of ontologies from existing standards and allows the creation of profiles and extensions of ontologies. We named this tool *RdfWiki* and explain it briefly in the next paragraph. Additionally, we need to create *ad-hoc* ontologies for areas of particular interest at regional level, and last but not least, to design and implement the infrastructure to facilitate creation, storing and querying of metadata for this volume of data.

RdfWiki Tool. DepMark worked with a single ontology: this fact presents an ideal setting for implementation, but lacks desirable properties such as composition and reusability of ontologies. We identified the necessity to develop a content management system driven by multiple ontologies. Projects with a similar aim are OntoWeb Portal [16] and ODESeW [17], but they are big-scale projects not aimed at small organizations.

Although RDF Schema (RDFS) can be used to specify ontologies, it presents several disadvantages in this case, particularly the weak datatype system and the lack of logical axioms to specify properties. That is why we are using OWL [18] (a language based on DAML+OIL [19] and proposed by the W3C as the ontology language for the Web) which supports the above requirements. For example, we can generate graphical user interfaces from ontologies taking a general model presenting a hierarchy of properties and classes (see Figure 6). Another useful feature of OWL is the possibility to model constraints.

The system is in the design phase. We called it *RdfWiki* to underline the designed features, such as public writable pages and rollback history. This application should not only permit the creation of metadata; Additionally, it may work like a collaborative space for creation of ontologies, by facilitating their composition from existing standards and allowing the creation of profiles and extensions of ontologies.

5 Conclusions

The success of RDF is based on the wide acceptance of a model. Experience has shown that information codifying web pages and information sources does not necessarily have to be located in the same place of that information. Moreover, it seems that building additional repositories is a more stable solution from the point of view of maintanability and updating,

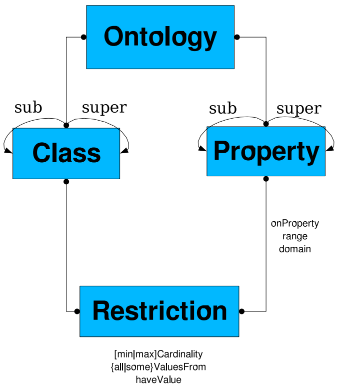


Fig. 6. Generic model of ontologies (compatible from RDF to OWL)

as well as application development.

The case of educational organizations shows that a model like the one developed in this paper has several advantages, not only from a technical point of view as argued in the paper, but also from a political and social point of view. In fact, the capture of information in an area so sensible, mutable and of public service as education, centralized solutions are neither convenient nor trustable by the community. To the best of our knowledge, there are no studies focusing on this social behaviour of people related to the educational area.

Our experience so far (with Depmark and in the design of RdfWiki) has highlighted some issues we consider relevant.

Markup must be made by people close to the data. It is not possible to bridge the semantic gap using only automatized extraction of metadata or centralized markup. Proliferation of online markup tools would make it possible to mark by hand in a distributed manner. Projects like this show that small scale markup efforts are a natural complement to big scale projects of massive automatic extraction of data from existent sources. The limitations of single-based ontology projects like Depmark makes it desirable to work with several ontologies. This direction is also closer to the spirit of the SW vision.

References

1. James Handler, Tim Berners-Lee, and Eric Miller. Integrating Applications on the Semantic Web. *Journal of the Institute of Electrical Engineers of Japan*, 122(10):676–680, October 2002.
2. Stefan Haustein and Jrg Pleumann. Easing Participation in the Semantic Web. In *Workshop at WWW2002, International Workshop on the Semantic Web*, May 2002.
3. David A. Wiley. *The Instructional Use of Learning Objects*, chapter Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. Agency for Instructional Technology and the Association for Educational Communications and Technology, 2000.
4. Sophie Lissonnet. A proposal for using metadata to support the building of an educational community . In *The Eleventh International World Wide Web Conference, WWW2002 Alternate Tracks*, 2002. May.

5. Ora Lassila and Ralph Swick. Resource Description Framework (RDF) Model and Syntax Specification. http://www.w3.org/ TR/ 1999/ REC-rdf-syntax-19990222/, February 1999.
6. Varun Ratnakar and Yolanda Gil. A Comparison of (Semantic) Markup Languages. *Proceedings of the 15th International FLAIRS Conference, Special Track on Semantic Web*, May 2002.
7. Dublin Core Metadata Initiative. http://purl.org/ dc.
8. Dan Brickley and R.V. Guha. RDF Vocabulary Description Language 1.0: RDF Schema. http://www.w3.org/ TR/ 2002/ WD-rdf-schema-20020430/, April 2002.
9. Brian McBride. Jena: Implementing the RDF Model and Syntax Specification. In Stefan Decker, Dieter Fensel, Amit Sheth, and Steffen Staab, editors, *Proceedings of the Second International Workshop on the Semantic Web - SemWeb'2001*, Hongkong, China, May 2001.
10. David Huynh, David Karger, and Dennis Quan. Haystack: A Platform for Creating, Organizing and Visualizing Information Using RDF. In *The Eleventh International World Wide Web Conference, WWW2002*, May 2002.
11. Ernesto Krsulovic and Claudio Gutirrez. Building Yearbooks with RDF. In A. Abraham et al., editor, *Soft Computing Systems: Design, Management and Applications*, pages 593–601. IOS Press, December 2002.
12. Ernesto Krsulovic and Claudio Gutierrez. Depmark, Marcado con metadatos de Departamentos Universitarios Chilenos. http://purl.org/ net/ depmark.
13. Canadian Heritage Information Network. Educational Metadata Standards. http://www.chin.gc.ca/ English/ Standards/ metadata_educational.html, April 2002.
14. European Schoolnet. European Treasury Browser. http://www.eun.org/ etb/.
15. Gail Hodge. Metadata Made Simpler. http://www.niso.org/ news/ Metadata_simpler.pdf, Annapolis Junction, MD, 2001.
16. Ontoweb. Ontoweb portal. http://www.ontoweb.org/.
17. O. Corcho, A. Gómez-Pérez, A. López-Cima, V. López-García, and M.C. Suárez-Figueroa. ODESeW. Automatic Generation of Knowledge Portals for Intranets and Extranets. http://webode.dia.fi.upm.es/sew/, 2003.
18. Mike Dean, Dan Connolly, Frank van Harmelen, James Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. OWL Web Ontology Language 1.0 Reference. http://www.w3.org/ TR/ owl-ref/, July 2002.
19. Ian Horrocks, Frank van Harmelen, and Peter Patel-Schneider. DAML+OIL. http://www.daml.org/ 2001/ 03/ daml+oil-index.html, March 2001.
20. *Computer Science Department Ontology*, v. 1.1, 2001.
http://www.cs.umd.edu/projects/plus/SHOE/onts/cs1.1.html
21. *International Workshop on Semantic Web and Databases*, http://swdb.semanticweb.org/