8th Intl. Protégé Conference. July 18-21, 2005. Madrid, Spain

Ontology-based browsing of bibliographic metadata

Joseba Abaitua, Josu Azpillaga*, JosuKa Díaz-Labrador, Jon Fernández, Inés Jacob, Txus Sánchez, Fernando Quintana

DELi Group, Universidad de Deusto, Bilbao, Spain <http://www.deli.deusto.es/>

*CodeSyntax, Eibar, Spain <http://www.codesyntax.com/>

Bibliographic metadata

A logical evolution in information management is one that scales up from the mere use of metadata to the ellaboration of a taxonomy and the later incorporation of an ontology. Whereas the transit from metadata into an structured taxonomy is a relatively easy and direct step, its mapping into an ontology is a very laborious process.

```
<inproceedings>
    <author>JosuKa Díaz Labrador and Joseba Abaitua Odriozola and Inés Jacob Taquet
    and Fernando Quintana Hernández and Garikoitz Araolaza</author>
    <title>Metadata for multilingual content management</title>
    <booktitle>Translating and the Computer 25. Conference Proceedings</booktitle>
    <year>2003</year>
</inproceedings>

<inbook>
    <author>Arantza Casillas and Joseba Abaitua and Raquel Martínez</author>
    <title>Recycling annotated parallel corpora for bilingual document
    composition</title>
    <pages>117-126</pages>
    <publisher>Springer</publisher>
    <year>2000</year>
</inbook>

<book>
    <author>Adolfo Arejita and Carmen Isasi and Itziar Túrrez</author>
    <title>Studia Philologica in Honorem Alfonso Irigoien</title>
    <publisher>Universidad de Deusto</publisher>
    <year>1998</year>
    <address>Bilbao</address>
</book>
```

BibTeX metadata for describing bibliographic information

Apart from the task of metadata mapping and the assignment of logical relations to the categories in the ontology, there is the problem of finding a suitable ontology. While in some domains well developed ontologies can be easily obtained (as with UNSPSC), in others no ontology is available (as in the case of XBRL taxonomies). Given these difficulties, some interesting strategies are being designed to synchronize both processes of creating metadata taxonomies and the ellaboration of the ontology (Handschuh and Staab, 2003).

@article: An article from a journal or magazine. Required fields: `author`, `title`, `journal`, `year`. Optional fields: `volume`, `number`, `pages`, `month`, `note`.
@book: A book with an explicit publisher. Required fields: `author` or `editor`, `title`, `publisher`, `year`. Optional fields: `volume` or `number`, `series`, `address`, `edition`, `month`, `note`.
@inbook: A part of a book, which may be a chapter (or section or whatever) and/or a range of pages. Required fields: `author` or `editor`, `title`, `chapter` and/or `pages`, `publisher`, `year`. Optional fields: `volume` or `number`, `series`, `type`, `address`, `edition`, `month`, `note`.
@incollection: A part of a book having its own title. Required fields: `author`, `title`, `booktitle`, `publisher`, `year`. Optional fields: `editor`, `volume` or `number`, `series`, `type`, `chapter`, `pages`, `address`, `edition`, `month`, `note`.
@inproceedings: An article in a conference proceedings. Required fields: `author`, `title`, `booktitle`, `year`. Optional fields: `editor`, `volume` or `number`, `series`, `pages`, `address`, `month`, `organization`, `publisher`, `note`.
...

BibTeX document taxonomy

```
<!ELEMENT bibtex:entry (
          bibtex:article | bibtex:book | bibtex:booklet | bibtex:manual |
          bibtex:techreport | bibtex:mastersthesis | bibtex:phdthesis |
          bibtex:inbook | bibtex:incollection | bibtex:proceedings |
          bibtex:inproceedings | bibtex:conference | bibtex:unpublished |
          bibtex:misc ) >

<!ENTITY  %   n.InProceedings "(
          bibtex:author, bibtex:title, bibtex:booktitle, bibtex:year,
          bibtex:editor?, (bibtex:volume | bibtex:number)?, bibtex:series?,
          bibtex:pages?, bibtex:address?, bibtex:month?,
          bibtex:organization?, bibtex:publisher?,
          bibtex:note?, %n.common;)">
```

BibTeX records and field types can be expressed through an XML Document Type Definition

A comprehensive corpus of bibliographic records have been tagged using BibTeX and Dublin Core metadata formats, which is an important advantage because ontologies for both metadata versions exist.

Discovering bibliographic records

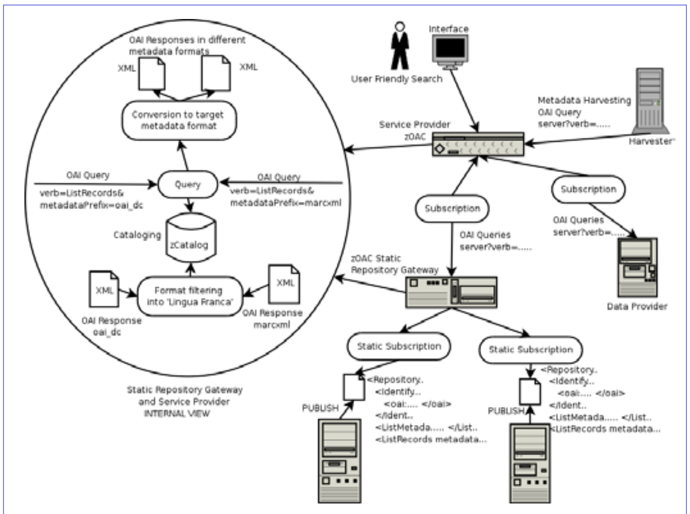
We lack a subject-based ontology (similar to Dolce or Cyc) that would facilitate the discovery process of these records (Welty, 1998). For the time being we do not know of any attempt to derive ontologies out of any the most widely used library classification schemes (LCC/LCSH, DDC, UDC, IFLA); although it should be a quite straitforward process. The main advantages of such schemes are that they improve thematic searches, harness multilingual access and increase interoperability with other services (Koch et al., 1997).

In any case, the methodological question that we will consider are the ways in which an ontology of structural metadata (such as BibTeX or Dublin Core metadata) can improve the discovery process of bibliographic information (this being the reverse process of the ontology-focused crwaling of web documentes presented by Ehrig and Maedche, 2003).

zOAC's architecture

The Open Archive Cataloger (OAC) project applies the OAI-PMH protocol for automatic metadata harvesting and aggregation of bibliographic records. It has been developed for the web application server ZOPE (http://zope.org), on top of the ZOpenArchives v1.2 product of the French company Pentila (http://www.pentila.com).

The main tools that have been developed are: the Static Repository Gateway, the Dynamic Data Provider, and the Service Provider.

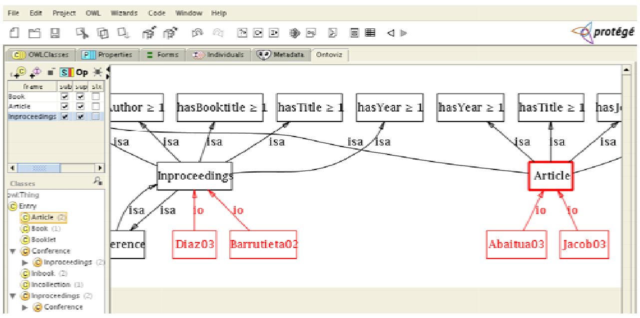


zOAC's architecture

Protégé for ontology-based metadata browsing

We are testing Protégé to evaluate the possibility of ontology-based metadata browsing. The challenge is important, taking into account the existence of powerful browsing engines solely based on the utilisation of metadata and string searching options, as those provided by Citeseer, Citebase or Google's Scholar services (Hitchcock et al., 2003).

```
konzeptu/ntxt/letrasUD_03.pdf</hasURL>

</Article>

<Incollection>
                <hasAuthor rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Carmen Isasi and
                     Ana Elejabeitia and
                     Iker Martínez and
                     Carmen Moral and
                     Santiago Pérez and
                     Joseba Pérez</hasAuthor>
                <hasTitle rdf:datatype="http://www.w3.org/2001/XMLSchema#string">La edición en el siglo XXI: nuevos
retos universitarios</hasTitle>
                <hasBooktitle rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Capacidades humanizadoras de
las TIC</hasBooktitle>
                <hasPublisher rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Universidad de Deusto</
hasPublisher>
                <hasYear rdf:datatype="http://www.w3.org/2001/XMLSchema#int">2004</hasYear>

                <hasEditor rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Inés Jacob</hasEditor>
                <hasISSN rdf:datatype="http://www.w3.org/2001/XMLSchema#string">84-7485-927-1</hasISSN>
                <hasLocation rdf:datatype="http://www.w3.org/2001/XMLSchema#string">pp. 381-388</hasLocation>
                <hasURL rdf:datatype="http://www.w3.org/2001/XMLSchema#string">http://www.deli.deusto.es/AboutUs/
Publications/JIS_04_SAI.pdf</hasURL>

            </Incollection>
  <owl:AnnotationProperty rdf:about="http://purl.org/dc/terms/bibliographicCitation">
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#DatatypeProperty"/>
  </owl:AnnotationProperty>
</rdf:RDF>

<!-- Created with Protege (with OWL Plugin 1.3, Build 225.4)  http://protege.stanford.edu -->
```



References

Marc Ehrig, Alexander Maedche. 2003. Ontology-focused crawling of web documents. In Proceedings of the 2003 ACM Symposium on Applied Computing (SAC-2003) (Gary Lamont, et al. editors). March 9-12, 2003, Melbourne, FL, USA. http://citeseer.ist.psu.edu/ehrig03ontologyfocused.html

Siegfried Handschuh, Steffen Staab. 2003. CREAM CREAting? Metadata for the Semantic Web. Computer Networks: The International Journal of Computer and Telecommunications Networking archive Volume 42 , Issue 5 (August 2003) table of contents Special issue: The Semantic Web: an evolution for a revolution Pages: 579 - 598. http://citeseer.ist.psu.edu/handschuh03cream.html; http://portal.acm.org/citation.cfm?id=873831

Steve Hitchcock, Arouna Woukeu, Tim Brody, Les Carr, Wendy Hall and Stevan Harnad. 2003 Evaluating Citebase, an open access Web-based citation-ranked search and impact discovery service. http://citeseer.ist.psu.edu/665664.html http://opcit.eprints.org/evaluation/Citebase-evaluation/evaluation-report-tr.html

Traugott Koch, Anna Brümmer, Debra Hiom, Marianne Peereboom, Alan Poulter, Emma Worsfold. The role of classification schemes in Internet resource description and discovery. Project deliverable D3.2. DESIRE - Development of a European Service for Information on Research and Education. http://www.ifla.org/documents/libraries/cataloging/classv10.pdf

Open Archive Cataloger. http://www.deli.deusto.es/OAC

Christopher A. Welty. The Ontological Nature of Subject Taxonomies. In Nicola Guarino Christopher A. Welty (eds). A Formal Ontology in Information Systems. Lecture Notes In Computer Science; Vol. 1937. http://www.cs.vassar.edu/faculty/welty/papers/fois-98/fois-98.ps