

Integration of SKOS and SRU in a distributed collaboration environment for archival material description.

Ricardo Eito-Brun
Universidad Carlos III de Madrid

Abstract

This poster describes the implementation of a solution based on the technical protocol SRU (*Search/Retrieve URL*) to search controlled vocabularies codified in SKOS. The application has been developed to access the engineering thesauri published by the Spanish Ministry of Civil Engineering - Ministerio de Fomento. SRU is a technical protocol designed by the Library of Congress to access remote bibliographic databases; it is the result of an effort to adapt the Z39.50 protocol to the World Wide Web. In this contribution, the author proposes a profile SRU to access remote thesauri from different software applications used to complete and edit metadata. Cataloguers can search these remote thesauri and assign descriptors to the records they create, regardless the metadata schema they are working with (EAD, MODS, MARCXML, etc.).

The proposed technical solution allows centres managing thesauri to expose and share their controlled vocabularies worldwide. This gives more visibility to these controlled vocabularies, and other libraries and archives may access them and benefit from the reuse of already-developed thesauri. The author considers that the capability of reusing these thesauri across a wide community of users is a key factor to demonstrate and justify the cost of the controlled vocabularies' development and maintenance.

Integration of SKOS and SRU in a distributed collaboration environment for archival material description.

Organization and representation of knowledge is a key activity in centres dedicated to the management of documentation. Libraries, archives, museums and documentation centres have systematically applied controlled vocabularies and classification schemas for the intellectual control and arrangement of materials and to make easier the access to their collections. The amount of materials that these centres collect has been increased by the addition of digital resources; this has resulted in a greater complexity, and in the need for working with additional metadata schemas for the description and management of materials and descriptive records. Metadata schemas like MODS, EAD, METS or PREMIS are now part of the standard professional practices of librarians and archivists.

Among knowledge organization techniques, SKOS (Simple Knowledge Organization System) is the most relevant proposal for encoding, transfer and exchange of thesauri. Pastor-Sanchez (2009) in mentions different initiatives developed before SKOS with the same objective of facilitating the encoding and exchange of controlled vocabularies: LIMBER (Language Independent Metadata Browsing of European Resources), CERES (California Environmental Resources Evaluation System), GEM (Gateway to Educational Materials), CALL (Center for Army Lessons Learned) Thesaurus, ETB (European Treasury Browser) and KAON/AGROVOC.

SKOS itself is a bridge between the current trends in web engineering and the traditional practices of libraries and archives for vocabulary control and the organization of indexing languages.

Published as a W3C recommendation in August 2009 by the Semantic Web Deployment Working

Group (the first draft was distributed in February 2008), SKOS offers a model to represent the structure and content of conceptual schemas like thesauri, classification systems, lists of subject headings or even taxonomies (Cantara, 2006). An important point to note is that SKOS should be considered not just as a tool to publish vocabularies, but also as a tool to represent the relationships between different conceptual schemas, as the specification gives the choice of establishing equivalences between concepts from different controlled vocabularies. The W3C recommendation acknowledges the need to take advantage of the experience of librarians and documentation specialists in knowledge organization in the development of the Semantic Web. Recommendations on how to use SKOS to encode different types of controlled vocabularies (thesauri, classification systems and taxonomies) are given by Miles and Perez-Agüera (2007).

SKOS features

SKOS is based on RDF. The concepts in the indexing language correspond to instances of one class, and the relationships between concepts and their descriptions are managed as declarations about those instances. One important feature of SKOS is that controlled vocabularies encoded in SKOS are not intended to represent a shared vision of reality (as happens with formal ontologies); the representation that we find in a SKOS-based vocabulary is restricted to the terms and descriptors taken from specific controlled vocabularies developed with a clear purpose and intended for a specific practical usage.

SKOS main characteristics may be summarized as follows:

- Concepts (defined as 'units of thought') are identified by means of URIs, and different labels, in one or more language, can be assigned to them.
- Concepts are grouped in 'concept schemas'.
- It is possible to add annotations to concepts.
- It is possible to create relationships between concepts, using the hierarchical and associative relationships used in indexing languages.

In addition, one feature especially relevant in our practical application of SKOS is the possibility of linking and relating concepts from different vocabularies. SKOS incorporates properties like `exactMatch` and `closeMatch` to indicate the different level of semantic similarity between concepts. Other properties like `broadMatch`, `narrowMatch` and `relatedMatch` may be used for those cases in which one concept has a meaning more or less generic or specific than another concept taken from a different vocabulary. The available options to establish equivalences between different vocabularies have been said to be insufficient by authors like McCulloch (2008) who carried out an analysis of the compatibility between different vocabularies (DDC, AAT, LCSH and MeSH) based on the different types of semantic correspondences proposed by Chaplan (1995).

SKOS does not indicate how concepts must be linked or related to the information resources to which they are assigned. We have the possibility of using SKOS descriptors from any metadata schema like Dublin Core, MODS or EAD, using the metadata and elements that these schemas incorporate to indicate the subject of the documents.

SKOS tools and projects

In the professional bibliography we can find the description of tools proposed to encode or migrate existing thesauri to SKOS (Pérez Agüera, 2004), (Ferreira, 2007), (Lacasta, 2007). The last one includes a wide list of tools developed for the creation and maintenance of thesauri (although not

all of them based on SKOS). In the area of medical and health documentation, Samwald (2008) proposed the usage of SKOS with the Bio-Zen ontology, and with other vocabularies like SIOC (Semantically Interlinked Online Communities), FOAF or Dublin Core. The author indicated the availability of relevant vocabularies in this area in SKOS format, like MeSH (Medical Subject Headings), Chemical Entities of Biological Interest or INOH Molecule Role Ontology.

Corey (2007) noted the role that documentation professionals can play in the evolution toward the Semantic Web by giving end-users a confidence in this initiative similar to the confidence they have in traditional information services. Contributing activities could be:

- a) Exposing collections and descriptive records to the Semantic Web;
- b) Migrating the controlled vocabularies and the semantic equivalences between them to Semantic Web formats (like SKOS);
- c) Sharing lessons learned.

One of the initiatives cited by Corey (2007) is a prototype service developed by OCLC – the OCLC GSAFD Vocabulary Service - to give users access to remote, controlled vocabularies from word processing tools like Microsoft Office. Other relevant initiatives where the mappings between different indexing languages are described are described by Angjeli (2008) and Nicholson (2006). The first one describes the STITCH project developed by the Bibliothèque National de France and Koninklijke Bibliotheek to analyse semantic interoperability when searching collections indexed with different controlled vocabularies. An updated review of this work is found in Angjeli, Isaac et al. (2009). The second one describes HiLT (High Level Thesaurus), a project investigating the use of SKOS combined with SRU to establish an M2M (Machine 2 Machine) system to search equivalences between concepts from different vocabularies, using the Dewey Decimal Classification as an intermediate language.

Coyle (2010) in a study about the use of RDF to encode bibliographic records based on Resource Description and Access (RDA) indicates the importance of SKOS and describes the initiatives of the Library of Congress and the National Library of Medicine to publish MeSH in SKOS.

An example of the possibility of extending SKOS to satisfy additional requirements, especially those related to the evolution of the terms in a thesauri, is provided by Tennis and Sutton (2008), who describe the inclusion of new labels to manage the 'descriptors lifecycle' as part of a Vocabulary Development Application developed for the National Science Digital Library Metadata Registry. These requirements – and the potential of extending SKOS to fulfill them – was also identified by Corey (2007). Prasad and Nabonita (2008) have proposed another SKOS extension for the annotation of resources based on facets.

Description of the proposed solution

This section summarizes the development of a software tool based on SKOS for accessing and searching thesauri developed by the Spanish Ministry of Civil Engineering (Ministerio de Fomento) from an EAD/ISAD(G) editor. Archivists cataloguing and describing archival materials (documents, photographs, manuscripts, etc.) are given the choice of creating and editing EAD/ISAD(G) finding aids as well as EAC-CPF authority records. When editing these finding aids, they can interact with remote SKOS repositories to locate and assign the descriptors they want to use as access points for their records.



Figure. 1. EAD/ISAD(G) Editor

The SKOS repository is a DBXML database. The interaction and communication between the EAD/ISAD(G) editor and the DBXML database is implemented by means of XML web services implemented on PHP, VBScript and compliant with the SRU technical protocol. Using the services and messages defined in this technical protocol, archivists cataloguing materials can select the controlled vocabulary they want to search, enter the terms, and restrict the search to different choices (for example just preferred terms, any term in the vocabulary, terms with a meaning more specific than the proposed one, etc). To this end, an SRU/CQL profile has been defined as part of this project.

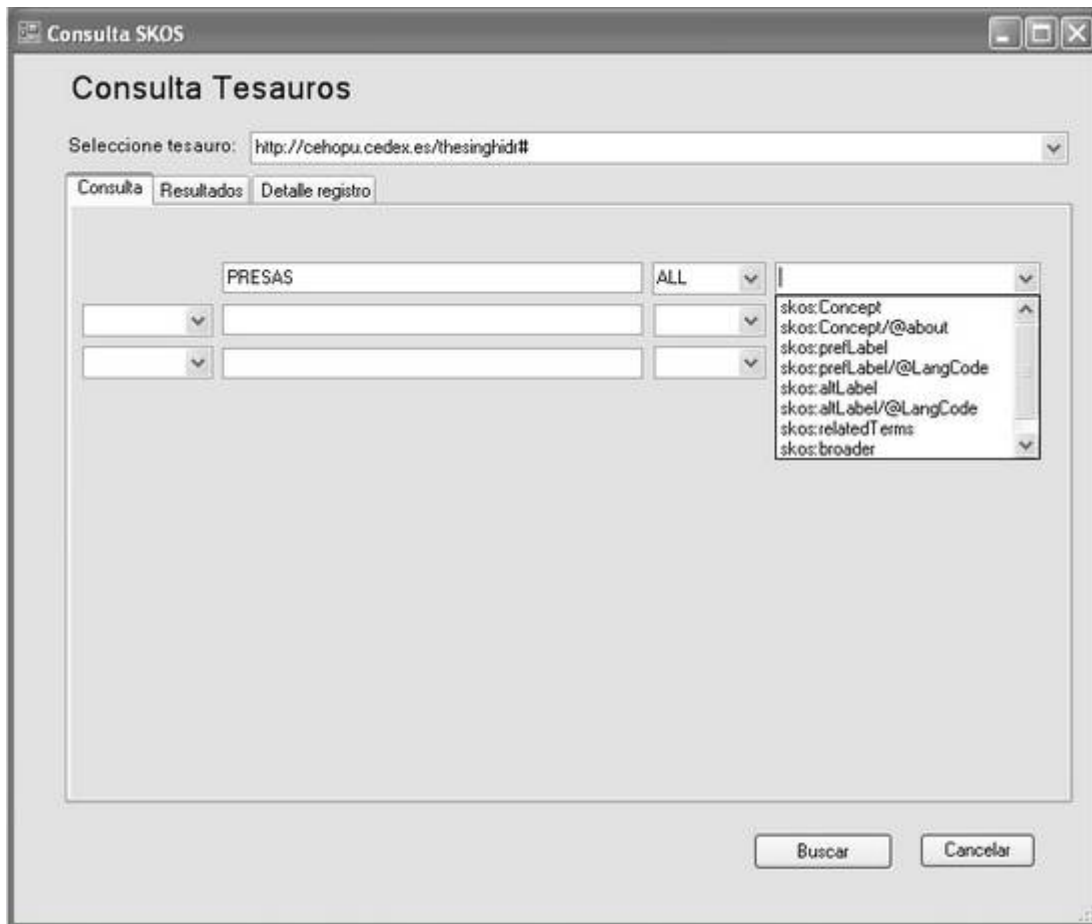


Figure 2. Searching the thesauri using SRU/CQL

Once the search criteria are entered, the local system (EAD editor) creates an SRU request that is directed to the database that corresponds to the chosen controlled vocabulary. After the execution of the search by the remote server, the client computer receives an SRU/XML response with the terms that meet the search criteria. The list of retrieved terms are shown to the archivists/cataloguer (preferred and non-preferred terms are displayed with different typography, and the system does not allow the assignment of non-preferred terms).

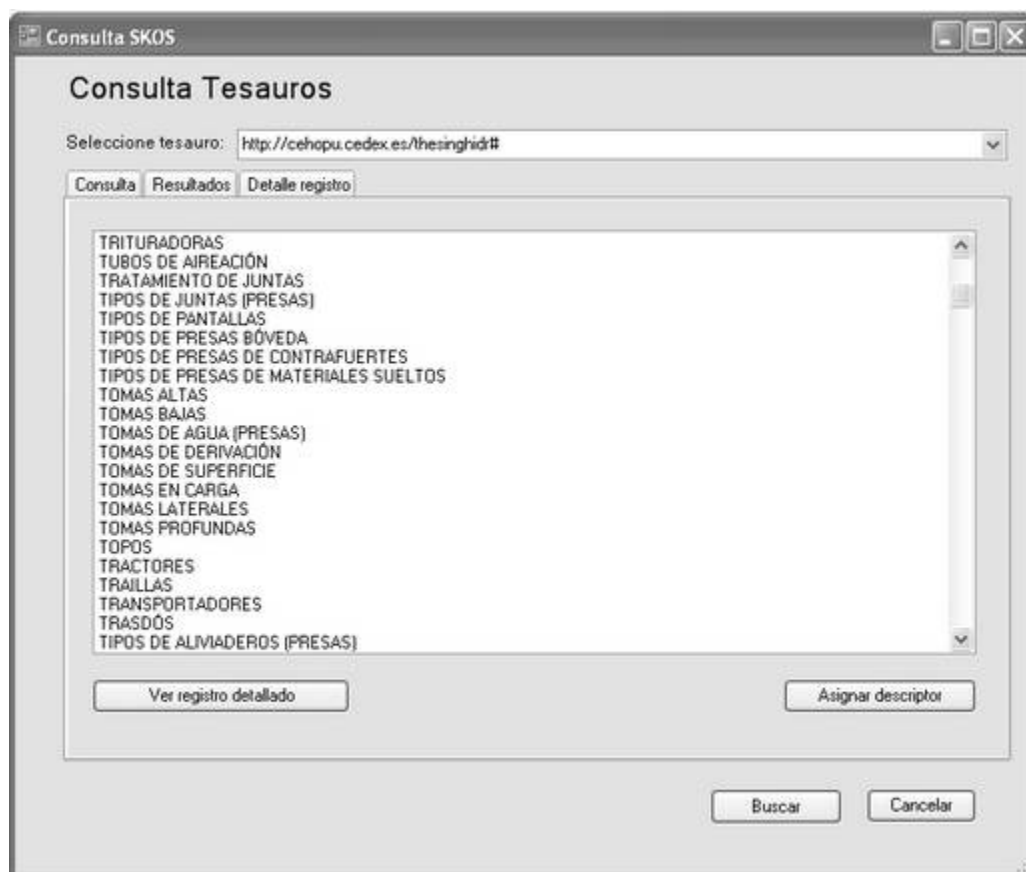


Figure 3. List of terms retrieved from the thesauri

From the list of results, the user can:

- Select one preferred term and assign it to the finding aid or IAD(G) description. The selected term shall be added to the EAD document in a new <subject> element. The attribute @source shall take as a value the URI of the controlled vocabulary. If the user is creating an authority record based on EAC-CPF, he/she may indicate to which EAC-CPF element the term must be assigned, and the source of the term shall be recorded in the @vocabularySource EAC-CPF attribute.
- Select one term from the list and request from the remote server the list of its related terms and the full information for this term (scope notes, etc.) available in the thesauri.

Regarding the selected method of storage for the SKOS thesauri within the DBXML database, each term in the controlled vocabulary – with all its relationships – is kept in a separate XML file. The full set of XML files (one per term) have been ingested in the DBXML database. This gives the choice of editing and updating each term, and its relationships, one by one, and improves the capability of managing semantic relationships between terms from different vocabularies.

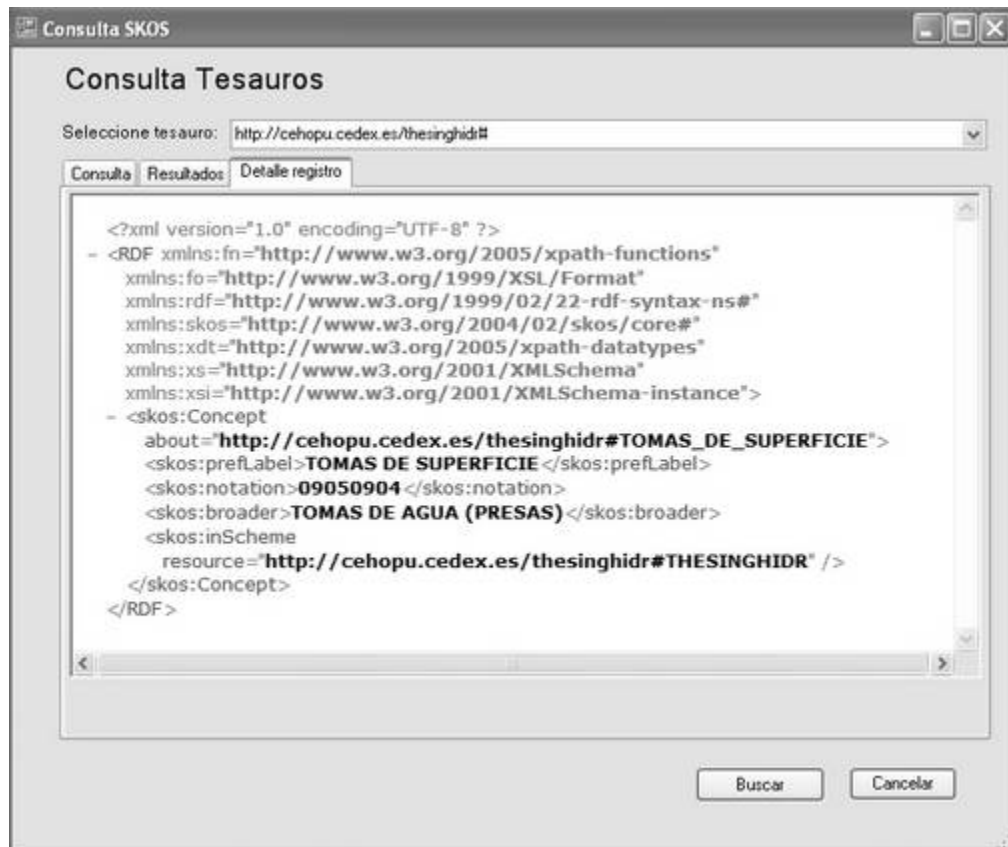


Figure 4. Storage of SKOS terms in the database

The interaction between the client application (the EAD/ISAD(G) editor) and the SKOS repository has been implemented by means of the messages defined in the SRU technical protocol. This is a sample request sent by the client to retrieve terms from the vocabulary:

```
http://server.es/sruSrvr/skos_processRequest.php?version=1.2&operation=searchRetrieve
&query=CONSULTA_CQL&maximumRecords=100&recordSchema=skos_summary
```

The response from the remote server will be also SRU-compliant, and will include all the terms / descriptors matching the search criteria. It is an XML message similar to this one:

```
<?xml version="1.0" encoding="UTF-8" ?>
<SRU:searchRetrieveResponse xmlns:SRU="http://www.loc.gov/zing/srw/">
<SRU:version>1.2</SRU:version>
<SRU:numberOfRecords>5</SRU:numberOfRecords>
<SRU:records xmlns:skos="http://www.w3c.org/2004/02/skos/core#">
  <SRU:record>
    <SRU:recordSchema>info:srw/schema/1/skos-v1.0</SRU:recordSchema>
    <SRU:recordPacking>xml</SRU:recordPacking>
    <SRU:recordData>
      <skos:Concept rdf:about="http://cehopu.cedex.es/thes#1002">
        <skos:prefLabel>Obras hidráulicas</skos:prefLabel>
      </skos:Concept>
    </SRU:recordData>
  </SRU:record>
</SRU:records>
</SRU:searchRetrieveResponse>
```

```

</SRU:record>
<SRU:record>
  <SRU:recordSchema>info:srw/schema/1/skos-v1.0</SRU:recordSchema>
  <SRU:recordPacking>xml</SRU:recordPacking>
  <SRU:recordData>
    <skos:Concept rdf:about="http://cehopu.cedex.es/thes#0350">
      <skos:altLabel>Drenaje hidráulico</skos:altLabel>
      <skos:prefLabel>Drenajes</skos:prefLabel>
    </skos:Concept>
  </SRU:recordData>
</SRU:record>
<!--REST OF THE RECORDS -->
</SRU:records>
</SRU:searchRetrieveResponse>

```

Conclusions

The development of the tool described above demonstrates the possibility of applying technical standards and protocols developed by the library community for bibliographic searches (SRU/CQL), in the context of archives for the purpose of accessing, and searching remote, controlled vocabularies based on SKOS. No similar experiences have been described in the professional literature regarding the application of SRU/CQL and SKOS in archives. Due to that, this project can be seen as an interesting starting point to demonstrate the feasibility of these standards and their potential for organizing and representing knowledge for the archival community.

In addition, the proposed tool gives developers of controlled vocabularies the choice of sharing and reusing their effort with other centres interested in using them. The usage of open standards like SRU/CQL and SKOS proposed in this work demonstrates the possibility of establishing collaborative networks for sharing, reusing and managing controlled vocabularies regardless of the tools and metadata they use to describe information resources.

Bibliography

- Angjeli, A. & Isaac, A. (2008). Semantic web and vocabularies interoperability: an experiment with illuminations collections. *IFLA Conference Proceedings, 2008*, 1-12.
- Angjeli, A., A. Isaac, et al. (2009). *Semantic Web and Vocabulary Interoperability: an Experiment with Illumination Collections*. *International Cataloging and Bibliographic Control*, 38(2) 25-29.
- Cantara, L. (2006). Encoding controlled vocabularies for the Semantic Web using SKOS Core. *OCLC Systems and Services*, 22(2) 111-114.
- Chaplan, M.A. (1995). Mapping laborline thesaurus to Library of Congress Subject headings, *Library Quarterly*, 65(1) 39-61.
- Corey, H. A. T., Barbara B. (2007). Library of Congress controlled vocabularies and their application to the Semantic Web. *Cataloging and Classification Quarterly*, 43(3) 47-68.
- Coyle, K. (2010). RDA in RDF. *Library Technology Reports*, 46(2) 26-36.
- Ferreira, D. (2007). *TemaTres: software libre para la gestión de tesauros*.
- Folch, H., B. Habert, et al. (2000). Navigable topic maps for overlaying multiple acquired semantic classifications. *Markup Languages: Theory and Practice*. 2, 269-280.

- García Jiménez, A. (2006). Una aproximació als llenguatges ¿documentals? en la web semàntica. *Item* (42) 33-50.
- García Torres, A., & Pradana-López, D. (2008). Reutilización de tesauros: el documentalista frente al reto de la web semántica. (Spanish). *Reusing thesauri: documentalists face the semantic web challenge*. (English) 17(1), 8-21.
- Gómez-Pérez, A. (2002). Ontology Specification Languages for the Semantic Web. *IEEE Intelligent Systems*, 17(1), 54-60.
- Lacasta, J., Nogueras-Iso, J., et al. (2007). ThManager: An Open Source Tool for Creating and Visualizing SKOS. *Information Technology and Libraries*, 26(3) 39-51.
- McCulloch, E. & Macgregor, G. (2008). Analysis of equivalence mapping for terminology services. *Journal of Information Science* 34(1), 70-92.
- Miles, A., & Perez-Agüera, J.R. (2007). SKOS: Simple Knowledge Organisation for the Web. *Cataloging and Classification Quarterly*, 43(3/4), 69-83.
- Nicholson, D., & McCulloch, E. (2005). Interoperable Subject Retrieval in a Distributed Multi-Scheme Environment: New Developments in the HILT Project. Retrieved from <http://cdlr.strath.ac.uk/pubs/nicholsond/ZaragosaPaperFinal.pdf>
- Pastor-Sanchez, J.-A., Mendez, J.M. et al. (2009). Advantages of thesaurus representation using the Simple Knowledge Organization System (SKOS) compared with proposed alternatives. *Information Research* 14(4), 19.
- Peis, E., & Herrera-Viedma, E. (2008). Modelo de servicio semántico de difusión selectiva de información (DSI) para bibliotecas digitales. (Spanish). *El Profesional de la Información*, 17(5): 519-525.
- Pérez Agüera, J. R. (2004). L'automatització de tesaurs i la seva utilització en el web semàntic. *BiD: textos universitaris de biblioteconomia y documentació* (13).
- Prasad, A. R. D., & Nabonita, G. (2008). Concept naming vs concept categorisation: a faceted approach to semantic annotation. *Online Information Review*, 32(4), 500-510.
- Salvador, S. (2006). Making use of upper ontologies to foster interoperability between SKOS concept schemes. *Online Information Review*, 30(3), 263-277.
- Samwald, M., & Adlassnig, K.P. (2008). The bio-zen plus ontology. *Applied Ontology*, 3, 213-217.
- Tennis, J. T., & Sutton, S. A. (2008). Extending the simple knowledge organization system for concept management in vocabulary development applications. *Journal of the American Society for Information Science and Technology*, 59(1), 25-37.

Corresponding author:

Ricardo Eito Brun can be contacted at reito@bib.uc3m.es