Use of NLM Medical Subject Headings with the MeSH2010 Thesaurus in the PORTAL-DOORS System

Carl TASWELL [1]
Global TeleGenetics, Inc., 8 Gilly Flower St., Ladera Ranch, CA 92694

Abstract. The NLM MeSH Thesaurus has been incorporated for use in the PORTAL-DOORS System (PDS) for resource metadata management on the semantic web. All 25588 descriptor records from the NLM 2010 MeSH Thesaurus have been exposed as web accessible resources by the PDS MeSH2010 Thesaurus implemented as a PDS PORTAL Registry operating as a RESTful web service. Examples of records from the PDS MeSH2010 PORTAL are demonstrated along with their use by records in other PDS PORTAL Registries that reference the concepts from the MeSH2010 Thesaurus. Use of this important biomedical terminology will greatly enhance the quality of metadata content of other PDS records thus improving cross-domain searches between different problem oriented domains and amongst different clinical specialty fields.

Keywords. Semantic Web, MeSH2010 Thesaurus, PORTAL-DOORS System.

1. Introduction

The PORTAL-DOORS System (PDS) for resource metadata management has been designed to address information retrieval problems caused by cybersilos, search engine oligopolies, the spread of misinformation, and continuing barriers to data interoperability in the transition from the original web to the semantic web and grid [6]. The architecture of PDS was modeled on the successful design of the IRIS-DNS System for the original web with hierarchically distributed mobile metadata [7]. The Internet Registry Information Service (IRIS) registers domain names while the Domain Name System (DNS) publishes domain addresses with mapping of names to addresses for the original web. Analogously, the Problem Oriented Registry of Tags And Labels (PORTAL) registers resource labels and tags while the Domain Ontology Oriented Resource System (DOORS) publishes resource locations and descriptions with mapping of labels to locations for the semantic web. This paper describes the most recent developments enabling enhanced description of resource metadata implemented for PDS as a result of the incorporation and use of the US NLM controlled vocabulary and thesaurus MeSH [1,4].

Facilities to enhance metadata description of resources entered in the PORTAL registries and DOORS directories of PDS are a necessary and important addition to improve

[1] Email: ctaswell@computer.org.

the content of each resource record. Incorporation and use of the MeSH 2010 Thesaurus has been prioritized as the first major controlled vocabulary to be integrated into PDS because of its important status and use by NLM for indexing of the medical literature. Currently, the MeSH Thesaurus is not published by NLM in a format that makes thesaurus concepts readily accessible as resolvable URIs with responses returned from a web service for integration with other tools and technologies of the various interpretations of the semantic web and grid. Moreover, the goal of exposing major vocabularies such as MeSH in a format that is exploitable by the PORTAL-DOORS System, the Linked Data initiative, or any other interpretation of semantic networks remains an important and necessary contribution to building the future semantic web and grid.

2. Methods and Results

An iterative process of software development and re-design has been pursued from the beginning of the project with PDS progressing through draft versions 0.1 to the current version 0.6. This iterative development has been maintained from a variety of perspectives including UML, SQL and XML modeling for PDS itself (the infrastructure system) as well as for the initial content managed by the system with the prototype biomedical registries GeneScene for genetics, ManRay for nuclear medicine, BrainWatch for brain imaging and neuropsychiatry, and BioPORT for biomedical computing [8].

All 25588 descriptor records from the NLM 2010 MeSH Thesaurus have been exposed as web accessible resources by the PDS MeSH2010 PORTAL operating as a RESTful web service [5]. Each descriptor record is published intact and unmodified from the original NLM source data by embedding it within the *other metadata* of the PDS *resource representation*. In addition to embedding the NLM descriptor record intact in the PDS record, several fields from each NLM descriptor record are also extracted and reproduced for each PDS record as other PDS fields such as the PDS *name* and *principal tag* to enable fast searching in the database. Each record published by the web service is referenceable via a PDS *resource label* so that it may also be used for metadata descriptions of other resources entered in the PORTAL registries and DOORS directories.

3. Use of the MeSH2010 PORTAL

The PORTAL-DOORS System specifies a set of data exchange interface requirements that facilitate interoperability and search across problem domains for both the original web and semantic web and grid [6]. Any PORTAL registry implemented for PDS may declare a set of constraints which define the focus of its problem scope as a *Problem Oriented Registry of Tags And Labels*. Resource representations entered as records for a given PORTAL registry should be validated against the set of constraints defined for the registry and expunged if not valid within the time period required by that registry [6].

For the MeSH2010 PORTAL introduced here, the problem oriented domain for the registry is declared simply as a thesaurus that reproduces the content of the US NLM MeSH 2010 Thesaurus in a manner and format interoperable and compliant with PDS. Thus, any entry in the PDS MeSH2010 Thesaurus must also be an entry in the US NLM MeSH 2010 Thesaurus. In this regard, the MeSH2010 PORTAL is closed to registration

of new resources other than administrative updates to match any updates at NLM in the source data. Serving as a thesaurus, the MeSH2010 PORTAL is thus different from the other prototype PORTAL registries (BioPORT, BrainWatch, GeneScene, ManRay) which are open for registration of new resources.

Public records in the MeSH2010 PORTAL are accessible via a RESTful web service available at `http://pds.portaldoors.net/mesh2010/` with server responses returning resource representations in XML format. Individual records can be retrieved simply by entering either the *canonical label* or *alias label* for the *resource representation*. For the first descriptor record in MeSH 2010 with DescriptorUI = "D000001", the corresponding PDS canonical label is

`http://pds.portaldoors.net/mesh2010/d000001`

and the PDS alias label is

`http://pds.portaldoors.net/mesh2010/calcimycin`

either of which will retrieve the same PDS resource representation. To demonstrate use of MeSH thesaurus concepts and records by other PDS records, the same example that first appeared as a pseudorecord in the virtual example in Section VII.A. of [6] is now implemented and available as a real record at the resolvable URL

`http://pds.biomedicalcomputing.net/bioport/elida`

This record was entered in the BioPORT Registry for which the problem oriented domain is declared as *biomedical computing* (see Section VIII of [6]).

4. Discussion

A PDS *resource representation* is only a representation of a resource, but not the resource itself. Resource representations stored in PORTAL registries and DOORS directories are only representations with metadata describing the resource. These representations refer to the resource but do not reproduce the resource. There are notable exceptions involving vocabularies such as the MeSH Thesaurus presented here for which each NLM MeSH descriptor record is reproduced intact and embedded within a PDS record. However, with regard to most other cases not involving vocabularies, recall the analogies for resource representations in PDS with the listings in a phone book and a library card catalogue as summarized in Table I of [7]. Neither the phone book nor the card catalogue reproduces the actual item described, instead only informing where that item is located and what kind of item it is.

Nevertheless, while maintaining interoperability with other components in PDS, any PORTAL registry requires some mechanism to limit registration of records only to those considered valid for the problem oriented domain declared as the scope for the particular registry. Prior to this report, the only validation mechanism implemented to date has been parsing the free form text for the presence of word stems in the *supporting tags* [8]. However, this paper introduces the use of *supporting labels* and an accompanying mechanism to test them for the presence of any requisite concept groups identified by entries in the PDS MeSH2010 Thesaurus as the corresponding implementation in PDS of the

NLM MeSH Thesaurus. With this new alternative approach to validating records for the problem oriented domain of each PORTAL, use of the MeSH Thesaurus in PDS should enable a more reliable scope declaration for each PORTAL in a manner consistent with the MeSH mission statement "to provide a reproducible partition of concepts relevant to biomedicine for purposes of organization of medical knowledge and information" [4].

While noting the distinction between the purpose of a database to store medical scientific data and the purpose of PDS registries and directories to use metadata to solve the data integration challenge for a given scientific problem, PDS also maintains the purpose of facilitating scientific social networking and semantic web linking (see Sections XI and XII of [6]). Although PORTAL registries may be declared private, all of the currently operating prototype registries are public and open to authored contributions, and the stated goal is to develop as many as possible that are public and open. These open public registries that allow contributions from a large number of investigators encourage active participation which in turn provides a better flow of suggestions for improvements to the official NLM MeSH Thesaurus.

5. Conclusion

Incorporation and use of the NLM MeSH controlled biomedical vocabulary and thesaurus to enhance the metadata description of resources entered within PDS should significantly improve the quality and utility of the content of PDS records for biomedical registries and applications including literature meta-analyses, clinical trials and medical imaging grids [3]. Continuing addition and integration of other biomedical terminologies including those encompassed by the UMLS metathesaurus [2] will further serve the PDS goal of interoperability for information retrieval and data integration.

References

[1] Medical Subject Headings (MeSH), US National Library of Medicine, 2010. URL `http://www.nlm.nih.gov/mesh/filelist.html`

[2] Unified Medical Language System (UMLS), US National Library of Medicine, 2010. URL `http://www.nlm.nih.gov/research/umls/index.html`

[3] Estrella, F., Hauer, T., McClatchey, R., Odeh, M., Rogulin, D., Solomonides, T.: Experiences of engineering Grid-based medical software., *International Journal of Medical Informatics*, **76**(8), Aug 2007, 621–632.

[4] Nelson, S.: Medical Terminologies That Work: The Example of MeSH, *Proceedings of I-SPAN 2009, The 10th International Symposium on Pervasive Systems, Algorithms and Networks*, IEEE Computer Society, December 2009, 380–384.

[5] Richardson, L., Ruby, S.: *RESTful Web Services*, O'Reilly Media, Inc., 2007.

[6] Taswell, C.: DOORS to the Semantic Web and Grid with a PORTAL for Biomedical Computing, *IEEE Transactions on Information Technology in Biomedicine*, **12**(2), Feb 2008, 191–204, In the Special Section on Bio-Grid.

[7] Taswell, C.: The Hierarchically Distributed Mobile Metadata (HDMM) Style of Architecture for Pervasive Metadata Networks, *Proceedings of I-SPAN 2009, The 10th International Symposium on Pervasive Systems, Algorithms and Networks*, IEEE Computer Society, December 2009, 315–320.

[8] Taswell, C.: Implementation of Prototype Biomedical Registries for PORTAL-DOORS, *Proceedings of the American Medical Informatics Association Summit on Translational Bioinformatics, San Francisco, CA*, Mar 2009, AMIA-0036-T2009.