Concept Validating Methods for Maintaining the Integrity of Problem Oriented Domains in the PORTAL-DOORS System

Carl Taswell
Global TeleGenetics
Ladera Ranch, CA, USA
ctaswell@computer.org

Abstract

As a distributed system of interacting PORTAL registries and DOORS directories, the PORTAL-DOORS System (PDS) provides management services for who-what-where metadata about both online and offline resources. PDS has been designed to facilitate search of varying scope both within and across registries and directories focused on different problem oriented domains. Maintaining the integrity of these problem oriented domains remains an essential requirement for maintaining the efficiency of search throughout the system. This report describes the new methods used in PDS to distinguish different specialty domains and demonstrates the approach for several registries including GeneScene and ManRay with concepts such as genes and radiopharmaceuticals. Metadata records are now tested by concept validating methods for the presence of any concepts required for each problem oriented domain. Invalid records are moved to a more appropriate registry or else deleted.

1 Introduction

The PORTAL-DOORS Project[1] (PDP) in biomedical informatics for the semantic web and grid began in 2006 with the goal of implementing a cyberinfrastructure capable of effective data integration, communication and interoperability across different specialty domains in the health care and life sciences. The origins of this project can be traced back to its progenitor, the GeneScene Source, which was initiated in 1999 and built as a directory of resources relevant to clinical genetics, genetic health care, and genetic sciences. Simple directories such as the original GeneScene Source did not anticipate the development of the semantic web, and thus, cannot benefit from interaction with the semantic web. This impediment to interoperability served as one of the important motivations for subsequent development of the PORTAL-DOORS System (PDS) [Taswell, 2008] which has since enabled reimplementation of the GeneScene Source Directory as the GeneScene PORTAL Registry.

Designed as a distributed online system, PDS is comprised of an interacting network of PORTAL registries and DOORS directories. The PORTAL servers operate as a resource label and tag registering system while the DOORS servers operate as a resource location and description publishing system. The names PORTAL and DOORS were derived respectively from the phrases *Problem Oriented Registries of Tags And Labels* and *Domain Ontology Oriented Resource System* that summarize their intended purposes.

Including GeneScene, there are now a total of 10 different specialty domain or problem oriented registries that are being used as prototypes for the ongoing development, implementation and revision of PDS. Both the original and revised design papers [Taswell, 2008; Taswell, 2010a] present details and discussion of the architecture for PDS within the context of comprehensive reviews of the literature. This report provides further details on the new methods used to maintain and distinguish different specialty domains in PDS using examples with two of the first registries, GeneScene and ManRay, and two of the most recent registries, HELPME and Osler.

2 Problem Oriented Domains

PDS specifies a set of data exchange interface requirements that facilitate interoperability and search across problem oriented domains for both the original web and semantic web [Taswell, 2008]. The administrators for any PORTAL registry implemented for PDS may declare a set of constraints which define the focus of its specialty domain or 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 that registry. If the representations are not validated for the registry within the time period required by that registry, the records considered invalid should either be deleted from the registry or else moved to a different more appropriate registry [Taswell, 2008]. Failure to do so, ie, failure to maintain the integrity of the domain scope for each registry by allowing irrelevant and/or inappropriate records to remain in any registry would defeat one of the most important purposes of building a problem oriented registry system.

The original PDS design [Taswell, 2008] introduced *supporting tags* (formatted as text phrases) while the revised PDS design [Taswell, 2010a] subsequently introduced *supporting labels* (formatted as URIs) for metadata records de-

[1] http://www.portaldoors.org

scribing resources. Supporting tags are intended for use with text phrases in a manner consistent with current conventional free-text tagging systems. Supporting labels are intended for use with URIs in a manner that references a controlled vocabulary, terminology or thesaurus as demonstrated in [Taswell, 2010b] for the NLM MeSH 2010 Thesaurus. All of the supporting tags and/or supporting labels for metadata records are marked as either *restricted* or *unrestricted* with regard to the registry's problem oriented constraints. If the tag or label is marked restricted, then it is subjected to validity testing for the restrictions imposed by the registry's constraints. If the tag or label is not marked restricted, then it is not validity tested. This approach enables each metadata record to be curated with some tags and labels that are validity tested as well as some that are not validity tested, thus permitting an author to provide as much metadata as desired while adhering to the restrictions required by the registry for compliance with its problem oriented domain.

3 Concept Validating Methods

All metadata records entered in a PORTAL registry are concept validity tested for compliance with any concept restrictions imposed by the scope definition declared by the administrators of the registry. For example, the GeneScene PORTAL Registry requires that any registered resource must maintain a metadata description with concepts relating to genetics, genes, DNA, RNA, etc, while the ManRay PORTAL Registry requires that resource records contain descriptions with concepts relating to radiopharmaceuticals, molecular imaging or nuclear medicine.

For the current prototype implementation of PDS registrars with draft version 0.7.1 of the PDS schema, several conventions have been adopted to facilitate initial entry of metadata records. The elements *entity name* and *entity nature* are considered special automatically restricted *supporting tags*, ie, supporting tags that are always automatically marked as restricted and thus always validity tested. Further, the algorithm tests in order first the *entity name* and *entity nature*, then any other *restricted supporting tags*, and last any *restricted supporting labels* terminating with successfull validation as soon as possible. In other words, if the name and nature are sufficient to validate the record successfully then the other tags and labels are not tested.

PDS employs a bootstrapping design with a self-referencing self-describing approach. Thus, the metadata record for a resource that is a PORTAL Registry itself contains the lists of constraints used to define the problem oriented domain for the registry. These lists can be found in the *registry restrictions* element of the *other metadata* element for metadata records available at

`http://pds.portaldoors.org/npds/portal`

for any of the registries selected by its name, for example, /genescene, /manray, /osler, and /helpme, each of which contains the word stems and phrases used for validity testing tags and the thesaurus concepts used for validity testing labels of other metadata records entered in that registry.

Note that an author or curator of a metadata record may choose an arbitrary number of either supporting tags and/or supporting labels to describe the resource entity. These tags and labels may or may not be related to the defining concepts that restrict the problem oriented domain for the PORTAL registry. To provide a more complete description of a resource entity, an author may choose to use some tags and labels that relate directly to the defining concepts for the PORTAL registry and some that do not. Thus, there is good cause for both restricted and unrestricted tags and labels. However, keep in mind that currently the concept validating methods for the records in the registry first evaluate the restricted supporting tags (including entity name and entity nature) and then the restricted supporting labels.

For the 4 registries presented above to demonstrate the use of restrictions to maintain the integrity of problem oriented domains, their scopes are declared essentially as *genetics* for GeneScene, *nuclear medicine* for ManRay, *personalized medicine* for Osler, and *Health Education Law Public Policy and Medical Ethics* for HELPME where detailed lists of word stems, word phrases, and thesaurus concepts can be reviewed by browsing the metadata record links above for each registry. These lists contain elements with the attributes *AndIndex* and *OrIndex* which correspond to the simple conjunctive and disjunctive Boolean logic that is used in the concept validating algorithm. The current algorithm validates records for presence of concepts. A future version may be enhanced with additional tests for absence of concepts.

4 Conclusion

New methods incorporating a concept validating algorithm have been implemented in PDS to maintain the integrity of problem oriented domains. Maintaining this integrity assures that searches within a given problem oriented registry will be as efficient as possible as a result of the absence of any records that are not relevant to the problem scope declared for the registry. Maintaining the integrity of each registry also assures that searches across a selected set of related registries will be as efficient as possible within the combined scope formed by the union of the scopes of the registries within the selected set.

References

[Taswell, 2008] Carl Taswell. DOORS to the semantic web and grid with a PORTAL for biomedical computing. *IEEE Transactions on Information Technology in Biomedicine*, 12(2):191–204, Feb 2008. In the Special Section on Bio-Grid.

[Taswell, 2010a] Carl Taswell. A distributed infrastructure for metadata about metadata: The HDMM architectural style and PORTAL-DOORS system. *Future Internet*, 2(2):156–189, 2010. In Special Issue on Metadata and Markup. Online at `http://www.mdpi.com/1999-5903/2/2/156/`.

[Taswell, 2010b] Carl Taswell. Use of NLM medical subject headings with the MeSH2010 thesaurus in the PORTAL-DOORS system. In Tony Solominides, Ignacio Blanquer, Vincent Breton, Tristan Glatard, and Yannick Legre, editors, *Proceedings of 8th HealthGrid 2010 Paris*, volume 159 of *Studies in Health Technology and Informatics*, pages 255–258. IOS Press, 2010.