Bridging the Semantic and Lexical Webs: Concept-Validating and Hypothesis-Exploring Ontologies for the Nexus-PORTAL-DOORS System

Adam CRAIG, Seung-Ho BAE and Carl TASWELL
Brain Health Alliance, Ladera Ranch, CA 92694, USA

Abstract

The Nexus-PORTAL-DOORS System (NPDS) has been designed with the Hierarchically Distributed Mobile Metadata (HDMM) architectural style to provide an infrastructure system for managing both lexical and semantic metadata about both virtual and physical entities. We describe here how compatibility between version 0.9 of the NPDS schema, the new NPDS-interfacing ontologies, and the domain-specific concept-validating hypothesis-exploring ontologies allows NPDS to bootstrap the semantic web onto the more developed lexical web. We then describe how this system will serve as the foundation of a planned platform for automated meta-analysis.

Keywords: Nexus-PORTAL-DOORS System, concept-validating ontology, hypothesis-exploring ontology, semantic web, message exchange, metadata management.

1. Nexus-PORTAL-DOORS System

The Nexus-PORTAL-DOORS System (NPDS) offers an approach by which individuals and organizations can manage their own repositories of semantic and lexical metadata about a problem domain of interest and share metadata records via a common message exchange format [1]. It consists of a metadata model schema, messaging specification, and network architecture for servers that distribute metadata about online and offline resources. NPDS includes four types of servers: The PORTAL registry serves as a lexical registry for uniquely identifying resources with representations that include URI identifiers, names, text descriptions, keyword tags, controlled vocabulary term labels, and cross-references. The DOORS directory provides a semantic directory for finding online and offline locations of identified resources via their URI identifiers and associated RDF descriptions. The Nexus diristry combines the functions of the PORTAL registry and DOORS directory into a single server. The NPDS components server maintains a collection of metadata records about NPDS services.

Analogous to the Internet Registry Information Service (IRIS) and Domain Name System (DNS) protocols, NPDS implements the Hierarchically Distributed Mobile Metadata (HDMM) architectural style to deliver accurate information with low latency [2]. Servers of the same type can form a hierarchy with primary servers that maintain master copies of records and that distribute them to secondary and caching servers. Components of the NPDS can communicate with each other and with web applications that retrieve records from them via a consistent RESTful web service API. The NPDS specification does not include the registrars through which human users or automated agents may register new resources with a PORTAL registry. We have created an example registrar service called Scribe that exposes a read-write RESTful web service API.

2. Prior Work

The first published description of the PORTAL-DOORS System introduced the PORTAL and DOORS servers in order to promote synergy between lexical and semantic metadata [1]. Version 0.6 introduced the Nexus server as a combined registry, directory, and registrar to simplify implementation and use by combining all functionality into a single server for simpler implementation [2]. Since then, we have worked with example implementations hosted at `http://npds.portaldoors.net/`, `http://npds.telegenetics.net/`, and `http://npds.brainhealthalliance.net/` to better understand the strengths and weaknesses of our approach in realistic usage scenarios [3], [4], [5]. In the previous iteration, version 0.8, we introduced a streamlined read-only RESTful NPDS API for PORTAL registries, DOORS directories, and Nexus diristires and a separate RESTful read-write API for Scribe registrars, simplifying development for client applications that only need to retrieve records [6].

By assigning each resource description a unique URI with resolvable URL from which it can be retrieved, using this and other URLs for cross-references between records, and delivering records using open web standards such as XML and JSON, NPDS offers a more concrete and structured realization of the ideals described in the linked-data principles [7]. We hope that NPDS as a distributed system will play a complementary role in the linked data ecosystem interfacing with those systems that operate as centralized repositories such as Wikidata [8] and domain-specific search of arbitrarily structured data such as DARPA's Memex projects [9]. In parallel with development of the NPDS specification itself, the example implementation of

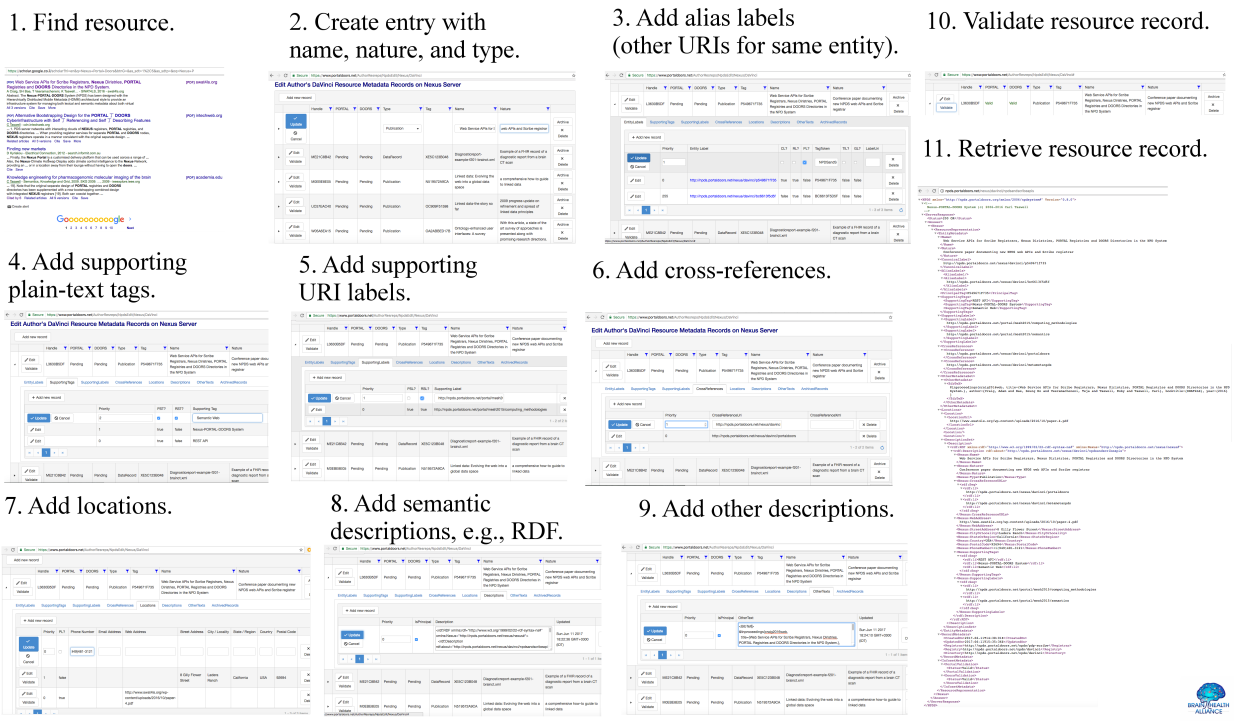

Figure 1: Example of typical work-flow in NPDS manual curation web application.

NPDS has followed a step-by-step development process in which each iteration has made fuller use of the system's potential:

v0.5 First production code implementation of partial PORTAL server functionality, partial DOORS server functionality, and AJAX-enabled web application for record curation. Version 0.5.4 was the last 0.5.* version published on 3/29/2009 [1].

v0.6 Implementation with ASP.net of RESTful web service API and AJAX-enabled curation application [2] (Figure 1).

v0.7 Implementation of lexical PORTAL and semantic DOORS functionality and interoperability with MeSH contolled vocabulary [3].

v0.8 Implementation of lexical-semantic Nexus diristries and streamlined RESTful read-only NPDS API and separate RESTful read-write Scribe API [6].

3. NPD Ontologies

The Nexus, PORTAL, and DOORS (NPD) ontologies bridge the gap between lexical metadata and semantic descriptions. Each server type is defined by an authoritative XML Schema Description (XSD) that describes the required fields each server must include in the records published through it and the optional fields for which NPDS client applications should check (Table 1). For each XML schema, we derived a corresponding formal ontology representing the elements of a Nexus, PORTAL, or DOORS record and the relationships among them in such a way that it is straight-forward to convert key lexical metadata fields into a semantic description of the resource. While the XML semantics reuse paradigm informed our choice of formalisms, we chose not to algorithmically convert every defined XSD type to a corresponding OWL class [10]. Instead, we selected the elements with information that would serve key functions for client applications, such as determining record provenance and currency and identifying different semantic descriptions on separate servers as describing alternate features of the same entity. Where appropriate, the NPD ontologies reference other widely used ontologies and map terms to equivalent terms in those ontologies. For example, the resource types in the Bibliographic Ontology (BIBO) are either equivalent to or subclasses of subclasses of the NPD resource type class, ET_Type based on the enumerated type of the same name in the XML schema [11].

4. Concept-Validating Ontologies

Large encyclopedic reference ontologies can be cumbersome for both users and developers. For users, especially those not versed in predicate logic, searching through a large tree of classes and examining their properties takes time and effort, especially when the ontology does not include the exact concept for which one is looking or a direct

relationship between two concepts of interest, thus requiring that the user cobble together a more complex query from the existing classes and properties. For developers, updating nodes in an ontology with hundreds or thousands of classes and properties may introduce inconsistencies that are difficult to resolve, making large ontologies more difficult to maintain. Organizing the encyclopedia of concepts into smaller, modular ontologies that cover the lexicons of special topics makes both issues more manageable. A user can focus on searching for classes and properties within a module of interest, and a developer can work on a module with less danger of introducing errors in other modules. This mode of thought has driven the development of the original ManRay ontology [12] and various other ontologies including the OBO Foundry ontologies, the Gene Ontology and Foundational Model of Anatomy, which are some of the most widely-used ontologies in the biomedical field [13].

Concurrently with the Nexus-PORTAL-DOORS System, Brain Health Alliance has been developing a collection of concept-validating ontologies founded on this same principle. Early examples include the ManRay ontology of nuclear medicine and diagnostic radiotracers [12] and the CTGaming ontology of clinical telegaming [4]. Each such ontology defines sufficient and necessary sets of concepts to which a knowledge resource must relate in order to fall within the scope of a particular problem domain. To add concept-validating constraints to an NPDS server, a domain expert takes URIs or plain text terms from the ontology and groups them into expressions in conjunctive normal form, e.g. (sensory OR language OR motor OR behavior) AND (onset) AND (neurodegenerative OR dementia) [5]. A record must feature at least one item from each group. A server can use multiple tests with different types of tags or labels, in which case the record need only pass one of the tests to be valid, ie, 'concept validated' for the registry [14] .

5. Hypothesis-Exploring Ontologies

More recently, Brain Health Alliance has taken this strategy a step further by introducing the hypothesis-exploring ontology, a compact ontology that includes only those concepts needed to describe the space of hypotheses regarding a well-defined scientific question. The initial example of the hypothesis-exploring ontology is the Sensory-Onset, Language-Onset, Motor-ONset (SOLOMON) ontology, which includes concepts needed to describe hypotheses on the relationships between onset type, disease progression, genes, brain regions, and proteinopathies in neurodegenerative diseases leading to cognitive decline and dementia [5]. As SOLOMON and ManRay illustrate, unlike OBO Foundry ontologies, NPDS-interfacing ontologies, whether only concept-validating or also hypothesis-exploring, may overlap with each other. The use of a canonical label as a unique identifier makes it clear whether two descriptions are of the same resource, even when they describe it using different ontologies. However, to avoid redundancy where it serves no purpose, we include as modules within the NPD ontologies the terms for shared use by all NPD concept-validating and hypothesis-exploring ontologies, such as 'Hypothesis', 'DependentVariable', and 'IndependentVariable', etc. We also map our use of terms to corresponding concepts where possible in the SWAN Biomedical Discourse Ontology [15].

6. Use-Case Example: Meta-Analysis

A Google Scholar search for "nucleus accumbens" returned 19,200 results published since 2011. Reading through the abstracts alone would take weeks or months. Furthermore, 70% of results from such a search are typically irrelevant, and 70% of relevant results missing [16]. The herculean task of sifting through these results often falls to the authors of literature review articles, valuable works that give readers new to a topic an overview of the current state of a problem domain. A particularly valuable type of literature review is the meta-analysis, which aggregates the effect sizes from numerous primary research articles testing the same hypothesis in order to achieve greater statistical power than any one primary study achieves alone [17]. The lengthy process of finding all the potentially relevant articles, selecting which ones are of sufficient novelty and quality to merit inclusion, and synthesizing their results into a coherent set of conclusions is often a difficult task that takes time away from the authors' own original research. In the field of cognitive neuroscience, the greater statistical power meta-analysis can achieve takes on special importance, because human subjects are both the most difficult to obtain in large number and the most variable in behavior. Although animal models can help to uncover the low-level mechanisms of neuronal function and even of collective behavior at the tissue scale, the human brain is unique in the animal kingdom, and an understanding of the relationship between genotype, gene expression, biochemistry, anatomy, and behavior in the human brain requires the study of human beings.

Brain Health Alliance is currently working to build a system for automated meta-analysis atop the NPDS infrastructure. In addition to the NPDS and Scribe services themselves, this system will consist of six components (Figure 2). Web applications enabling curation of metadata provide a human-friendly user interface for the Scribe registrars. Focused web crawlers retrieve information about knowledge resources relevant to a problem domain from databases, search engines, and other resources [18]. Such programs can aid human curators in populating NPDS servers with metadata. Natural language processors translate natural language questions into SPARQL queries and output from statistical analysis packages into natural language answers [19]. Hypothesis-exploring ontologies allow for more direct translation of questions from domain

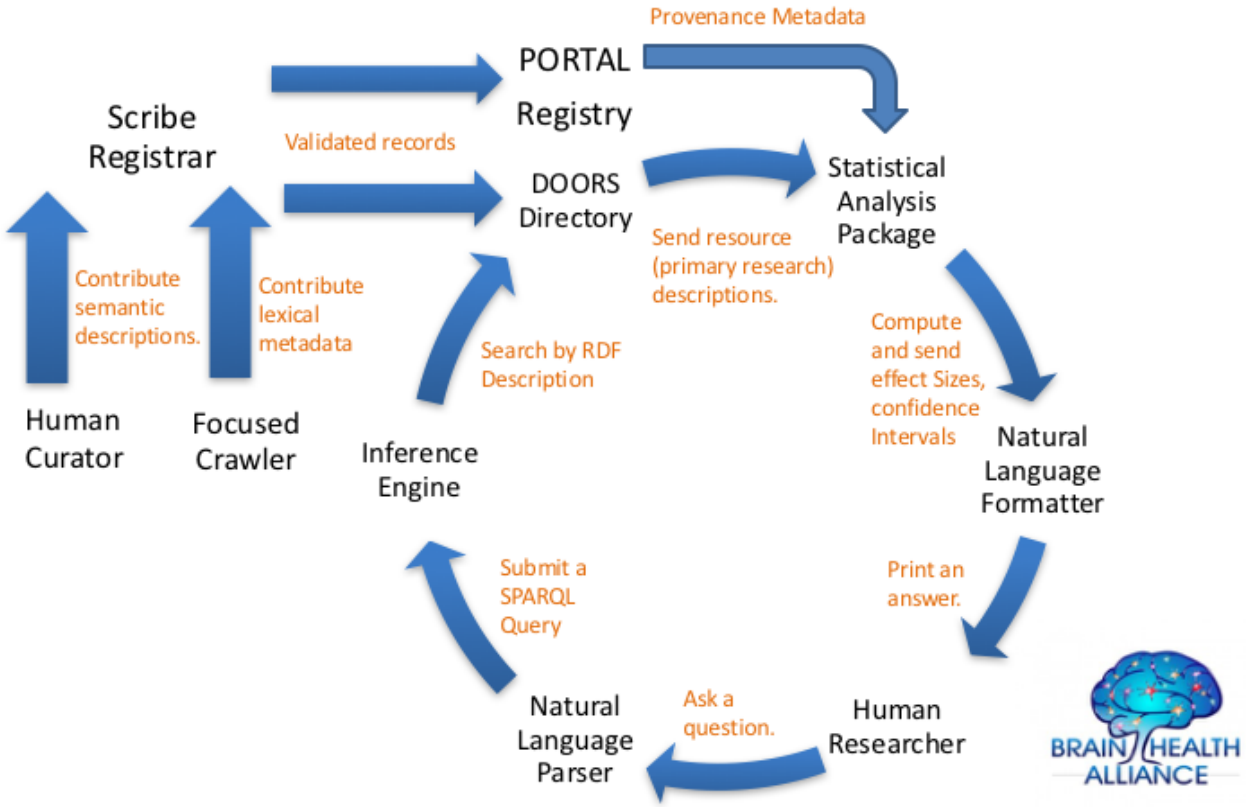

Figure 2: Flow of information between components in the proposed system for automated meta-analysis

experts into SPARQL queries by providing a compact set of the most relevant concepts and relationships referencing more comprehensive foundational and domain ontologies to allow for query expansion [5]. Inference engines will expand queries and extract the relevant information from semantic descriptions of resources in DOORS or Nexus records [20]. Statistical analysis packages will compute aggregate effect sizes and confidence intervals from the retrieved data [21].

7. Conclusion

Similar to previous versions of the Nexus-PORTAL-DOORS System, NPDS v0.9 provides an infrastructure system for the semantic web hybridized with the lexical web. This current iteration brings NPDS another step closer to bridging successfully the gap between the lexical and semantic webs by interfacing with the NPD ontologies, which in turn interface with domain-specific concept-validating hypothesis-exploring ontologies. These compact and easily updated ontologies will be instrumental in developing applications that perform semantic search for resources registered with NPDS, including applications for automated meta-analysis.

References

[1] C. Taswell, "DOORS to the semantic web and grid with a PORTAL for biomedical computing," **IEEE Trans Inf Technol Biomed**, vol. 12, no. 2, pp. 191–204, 2 Mar. 2008, In the Special Section on BioGrid, ISSN: 1089-7771. DOI: 10.1109/titb.2007.905861.

[2] ——, "A distributed infrastructure for metadata about metadata: The HDMM architectural style and PORTAL-DOORS system," **Future Internet**, vol. 2, no. 2, pp. 156–189, 2010, In Special Issue on Metadata and Markup., ISSN: 1999-5903. DOI: 10.3390/fi2020156. [Online]. Available: http://www.mdpi.com/1999-5903/2/2/156/.

[3] ——, "Use of the MeSH thesaurus in the PORTAL-DOORS system," in **Proceedings AMIA 2010 Symposium on Clinical Research Informatics, San Francisco CA**, 2010, AMIA-033-C2010.

[4] ——, "A new PDS PORTAL for clinical telegaming rehabilitation and intervention," in **Proceedings of 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM10)**, 2010.

[5] M. Skarzynski, A. Craig, and C. Taswell, "SOLOMON: An ontology for sensory-onset, language-onset and motor-onset dementias," in **Bioinformatics and Biomedicine (BIBM), 2015**

IEEE International Conference on, IEEE, 2015, pp. 969–972.

[6] A. Craig, S.-H. Bae, T. Veeramacheneni, *et al.*, "Web service APIs for Scribe registrars, Nexus diristries, PORTAL registries and DOORS directories in the NPD system," in **Proceedings of the 9th International SWAT4LS Conference**, Amsterdam, Netherlands, 2016. [Online]. Available: http://ceur-ws.org/Vol-1795/paper4.pdf.

[7] T. Berners-Lee, *Linked data*, 2006.

[8] E. Mitraka, A. Waagmeester, S. Burgstaller-Muehlbacher, *et al.*, "Wikidata: A platform for data integration and dissemination for the life sciences and beyond," **bioRxiv**, p. 031 971, 2015.

[9] B. Wilson, L. McGibbney, C. Mattmann, *et al.*, "Memexgate: Unearthing latent content features for improved search and relevancy ranking across scientific literature," in **AGU Fall Meeting Abstracts**, 2015.

[10] R. G. González, **A semantic web approach to digital rights management**. Universitat Pompeu Fabra, 2007.

[11] B. D'Arcus and F. Giasson, *Bibliographic ontology specification*, 2009.

[12] C. Taswell, B. Franc, and R. Hawkins, "The ManRay project: Initial development of a web-enabled ontology for nuclear medicine," in **Proceedings of the 53rd Annual Meeting of the Society of Nuclear Medicine, San Diego, CA**, Jun. 2006, p. 1431.

[13] B. Smith, M. Ashburner, C. Rosse, *et al.*, "The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration.," **Nat Biotechnol**, vol. 25, no. 11, pp. 1251–1255, 2007.

[14] C. Taswell, "Concept validating methods for maintaining the integrity of problem oriented domains in the PORTAL-DOORS system," in **IDAMAP 2010: Intelligent Data Analysis in Biomedicine and Pharmacology**, S. Swift and K. Phillips, Eds.

[15] P. Ciccarese, E. Wu, G. Wong, *et al.*, "The SWAN biomedical discourse ontology," **Journal of Biomedical Informatics**, vol. 41, no. 5, pp. 739–751, 2008, ISSN: 1532-0464.

[16] A. Rzhetsky, M. Seringhaus, and M. Gerstein, "Seeking a new biology through text mining," **Cell**, vol. 134, no. 1, pp. 9–13, 2008, ISSN: 0092-8674.

[17] H. Cooper, L. V. Hedges, and J. C. Valentine, **The handbook of research synthesis and meta-analysis**. Russell Sage Foundation, 2009.

[18] S. Chakrabarti, M. Van den Berg, and B. Dom, "Focused crawling: A new approach to topic-specific web resource discovery," **Computer networks**, vol. 31, no. 11, pp. 1623–1640, 1999.

[19] E. Kaufmann and A. Bernstein, "How useful are natural language interfaces to the semantic web for casual end-users?" In *The Semantic Web*, Springer, 2007, pp. 281–294.

[20] J. Mayfield and T. Finin, "Information retrieval on the semantic web: Integrating inference and retrieval," in **Proceedings of the SIGIR Workshop on the Semantic Web**, 2003.

[21] R. J. Grissom and J. J. Kim, **Effect sizes for research: Univariate and multivariate applications**. Routledge, 2012.

Schema/Ontology Class	NPDS Message Element	side	level	req.	Function
CT_ResourceRepresentationPortal	ResourceRepresentation	PORTAL	N/A	yes	lexical metadata except location
CT_ResourceRepresentationDoors	ResourceRepresentation	DOORS	N/A	yes	location and semantic description
CT_ResourceRepresentationNexus	ResourceRepresentation	Nexus	N/A	yes	both lexical and semantic information
CT_EntityLevel1Metadata	EntityMetadata	PORTAL	entity	yes	description of the resource
CT_Name	Name	all	entity	no	plain text name
CT_Nature	Nature	all	entity	no	plain text description
CT_EntityCanonicalLabel	CanonicalLabel	all	entity	yes	main NPDS URI
CT_EntityAliasLabel_Set	AliasLabels	PORTAL	entity	no	alternate NPDS URIs
ST_PrincipalTag	PrincipalTag	PORTAL	entity	no	plain-text unique identifier
CT_SupportingTag_Set	SupportingTags	PORTAL	entity	no	relevant plain-text tags
CT_SupportingLabel_Set	SupportingLabels	PORTAL	entity	no	relevant controlled vocabulary IRIs
CT_CrossReference_Set	CrossReferences	PORTAL	entity	no	related NPDS record URIs
CT_OtherEntityLabel	OtherEntity	PORTAL	entity	no	any non-NPDS IRI identifier
CT_OtherMetadata_Set	OtherMetadata	PORTAL	entity	no	any other metadata
CT_ContactLabel	Contact	PORTAL	entity	no	IRI of person to contact about entity
CT_OwnerLabel	Owner	PORTAL	entity	no	IRI of entity owner
CT_Location_Set	Locations	DOORS	entity	yes	physical or on-line locations of entity
CT_Description_Set	Descriptions	DOORS	entity	no	set of RDF descriptions
CT_RecordLevel2Metadata	ResourceRepresentation	PORTAL	record	yes	description of the NPD record
CT_PersonLabel	CreatedBy	all	record	no	IRI of record creator
xsd:dateTime	CreatedOn	all	record	no	date-time record created
CT_PersonLabel	UpdatedBy	all	record	no	IRI of last updating user
xsd:dateTime	UpdatedOn	all	record	no	date-time last updated
CT_PersonLabel	ManagedBy	all	record	no	IRI of user managing the record
CT_DirectoryLabel	Directory	PORTAL	record	no	DOORS URL
CT_RegistryLabel	Registry	DOORS	record	no	PORTAL URL
CT_InfosetLevel3MetadataPortal	InfosetMetadata	PORTAL	infoset	yes	record validation status
CT_InfosetValidation	PortalValidation	PORTAL	infoset	yes	validation vs PORTAL constraints
CT_InfosetValidation	DoorsValidation	DOORS	infoset	yes	validation vs DOORS constraints

Table 1: Key elements of Nexus, PORTAL, and DOORS resource representations