**Open Repositories 13 proposal**

Title: WordShack: A Vocabulary Registry for Preservation Repositories

Authors: Robin Wendler, Bobbi Fox, Andrea Goethals, Chris Vicary, Wendy Gogel, Randy Stern

Keywords: preservation, PREMIS, vocabulary, MADS, metadata, authority, schema

Track: General conference

Type: “24x7”

Abstract:

Digital library systems often require controlled vocabularies. They facilitate efficient collection management and improve discovery of content by end users by providing more comprehensive and more precise search results. Unfortunately, these vocabulary lists are often duplicated and managed individually in local databases or applications, leading to duplication of development efforts and the use of similar but different terms to represent the same concept. Recognizing these problems, in 2009, Harvard Library began designing a central vocabulary registry, “WordShack”, for use across the Library’s digital library applications and systems, including its digital preservation repository, the Digital Repository Service (DRS).

WordShack is a registry for controlled vocabulary terms used in the Harvard Library’s digital preservation suite of services.  Currently integrated with the DRS and Email Archiving Service beta, WordShack provides

* Persistent identifiers for vocabulary concepts
* An authoritative source for the current preferred term for a concept
* References from alternate terms for a concept
* Single maintenance of vocabulary concepts shared across systems
* Persistent storage of information needed in Harvard’s preservation systems that would be dropped from external sources in the course of normal business (e.g., names of Harvard agents after they leave the university.)

The need for controlled vocabulary is especially evident within the metadata managed within the DRS. For example, in the PREMIS schema for preservation metadata, significant events typically include the name of a person, organization or software program associated with the event. In order to eliminate ambiguity about the entity represented by a name, a vocabulary control mechanism such as WordShack is desirable. By representing these entities in the metadata as persistent identifiers resolving to authority-controlled names within WordShack, and not solely by the name strings themselves, the agents related to events can be unambiguously identified and managed centrally.

In addition to PREMIS event agents, WordShack-managed terms have been found to be useful for many other metadata fields stored in the DRS. A few examples:

* Email addresses and associated persons come into play as metadata associated with archived email collections. By controlling these terms, curators and archivists can unambiguously document the collection source or content creators, as well as tie the email collections to other content associated with the same people.
* Topics are subject terms that can be used to characterize collections of digital objects. By using controlled terms for these topics, it will make it easier for end users to discover the content, from any of the Library’s applications.

The basic unit within WordShack is a “term”. In its simplest form, a term has a unique ID, a preferred value and zero or more variant values. The term types that are supported include topic, software, administrative category, administrative flag, email address, person and organization. Each term type has associated metadata stored and managed in WordShack. For example, a software term includes a name, version and genre. The schema permits one to indicate the authority source of the term, for example to indicate that the term was exported from a different vocabulary registry. Provision is made for including the “foreign” vocabulary’s URI or other formal identifier.

The WordShack implementation includes a database, an API, and user interface widgets. The API is a RESTful API for CRUD (create, read, update, and delete) operations. The API also supports CRUD operations on relationships between terms – for example, relating an email address to a person – as well as term deprecation, reactivation, and superseding operations. Throughout, the API supports variant as well as preferred names for terms. Underlying the API is a relational database modeled on the Library of Congress’ MADS schema.

WordShack also includes a set of jQuery UI JavaScript user interface widgets for embedding controlled vocabulary within administrative and discovery web applications. “Select” widgets provide an auto-complete drop down user interface for locating terms by typing character strings embedded in any term variant. “Edit” widgets provide the ability to create or modify all of the metadata elements associated with a term. Use of the widgets is of course optional – new user interface access methods, such as pull down lists, checkboxes, etc. can be programmed directly using the WordShack API. The WordShack widgets and API have been used in creating an administrative interface for managing WordShack terms and for associating terms with DRS objects, as well as for specifying terms in the DRS ingest application and an email archiving curation interface. The presentation will include screen shots of this user interface.

If there is sufficient interest from the repository community, WordShack will be made available as open source software available on a public code repository. We welcome contributors who might extend the core code, for example by creating an RDF export/import facility, or for enabling federation and interoperation with other authority control systems.