Persistent Identifiers for Metadata Terms in a Crowd-Sourced Vocabulary

John Kunze California Digital Library USA jak@ucop.edu

Greg Janee UC Santa Barbara USA gjanee@ucop.edu

Christopher Patton UC Davis USA cjpatton@ucdavis.edu

Abstract

Unique, persistent identifiers for vocabulary term concepts are critical for metadata (DC¹, SKOS², etc). This comes as no surprise to followers of Linked Data³, for whom this first principle of the semantic web is a *sine qua non* for automatic reasoning with web content. It is even more important to metadata users who need a precise way to reference a particular concept when the term may have more than one definition.

Such is the case for the Sealce Metadictionary⁴, a crowd-sourced online dictionary of metadata terms in which multiple competing definitions are expected to be common and to co-exist indefinitely. Anyone can register and login in order to create new terms, edit their own terms, and comment and vote on others’ terms. Typical use will be that someone, without logging in, searches for and inserts terms they find into metadata that they’re creating to describe their own research. If unsatisfied with the terms that they found – or didn’t find – they can login and take action, which means anything from up- and down-voting terms, commenting on others’ terms, or adding and editing their own terms. Typical users will be research scientists trying to describe their datasets.

At the moment, Sealce’s internally generated database table row numbers are the only unique identifiers to disambiguate term concepts. As these numbers are not globally unique and not reproducible if the database is ever reloaded in a different order, we envisage creating another table column and populating it with stable identifiers from a recognized naming authority. To populate this column, we plan to extend Sealce to mint identifiers from automatically from EZID⁵ at the University of California, an established identifier service known for its open, identifier-scheme-agnostic architecture in support of research datasets and cultural heritage material.

In this presentation, we will discuss and seek feedback on a number of questions arising from the application of persistent identifiers (pids) to metadata elements. Questions include:

- Does it make sense to assign pids to element values as well as to element names?
- Should term relationships have pids? E.g., broader, narrower, synonym, antonym, etc.
- Research data relies on units, types, and other sub-vocabularies. Should they get pids?
- What should the experience of resolving a term pid be?

Participating in CAMP-4-DATA with attendees from diverse disciplines will provide valuable feedback for our development plan.

---

² *Simple Knowledge Organization System*
⁴ A product of the Preservation and Metadata Working Group within the NSF-funded DataONE project.
⁵ UC Curation Center. *Long-term identifiers made easy* (n2t.net/ezid)