Collaborative Building of Controlled Vocabulary Crosswalks

Mateusz Kaczmarek, Sebastian Ryszard Kruk, Adam Gzella

ABSTRACT: This paper presents analyses how to build an algorithm for building controlled vocabulary crosswalks, using JOnto framework for annotating resources with KOS. We will present how this framework can be used in a context of JeromeDL – a Semantic Digital Libraries and Social Semantic Collaborative Filtering.

One of the main features of classic libraries is metadata, which also is the key aspect of the Semantic Web. Librarians in the process of resources annotation use different kinds of Knowledge Organization Systems; KOS range from controlled vocabularies to classifications and categories (e.g., taxonomies) and to relationship lists (e.g., thesauri). The diversity of controlled vocabularies, used by various libraries and organizations, became a bottleneck for efficient information exchange between different entities. Even though a simple one-to-one mapping could be established, based on the similarities between names of concepts, we cannot derive information about the hierarchy between concepts from two different KOS.

One of the solutions to this problem is to create an algorithm based on data delivered by large community of users using many classification schemata at once. The rationale behind it is that similar resources can be described by equivalent concepts taken from different taxonomies. The more annotations are collected, the more precise the result of this crosswalk will be.

We are going to present JOnto, a component that allows to annotate resources with taxonomy and thesauri entries. At the moment it comes with the pre-installed support for WordNet, Dmoz, Dewey Decimal Classification, and Polish Thematic Classification; it can be, however, easily extended to support other classification schemata. It has been successfully implemented in JeromeDL, and in Social Semantic Collaborative Filtering (SSCF). In the first case it allows librarians to used different classification schemata, and WordNet thesauri, to classify and annotate resources in the library.

In the latter case, SSCF is a component primary delivered for JeromeDL. It allows a community of users to annotate and share resources. Each user maintains his/her own classification through a taxonomy of bookmark folders. Users can share these folders among each other; in a result, the whole community maintains a list of terms (bookmark folders), which is instantiated as classification taxonomy for each user. SSCF uses JOnto to help users to ground each term with existing KOS.

In both cases, community of librarians and users, deliver some hints about possible mappings between concepts from different controlled vocabularies. In our presentation, we would like to discuss various algorithms and approaches to build crosswalks between different classification schemata, based on data aggregated with JOnto. We will present how these algorithms can improve:

- Annotation process by suggesting, and automatically pre-selecting, concepts from different taxonomies
- Search and retrieval process in heterogeneous networks of digital libraries.

We will also discuss quality metrics that can be applied to evaluate our solution.