Using SKOS and FOAF for Name Authority in the Institutional Repository

Tom Johnson, Digital Applications Librarian, Oregon State University

Submission Type: Conference Paper/Presentation

Abstract:

Name ambiguity is a widespread problem in Institutional Repository metadata. A lack of authority control, combined with often user-submitted metadata, leads to individuals and organizations appearing in a variety of name permutations in a single system. Equally damaging is namespace collision—distinct entities referred to by a single name. While repository software has improved in its ability to utilize externally maintained name authority databases, the range of names used by repositories often expands beyond what is controlled by traditional library sources.

Using SKOS and FOAF, Oregon State University Libraries is creating minimal but extensible Linked Data name authority records for use with our institutional repository. These records can be locally maintained and are interoperable with other Linked Data authority systems, such as id.loc.gov and VIAF. This paper details the approach to authority control we are establishing based on our Linked Data model and the challenges we have faced in disambiguating existing metadata.

Extended Proposal:

The full paper and presentation will provide a synthesis of topics, addressing the full range of issues involved in establishing a local authority system . Our goal is to provide enough detail that other institutions can reproduce our work. To that end, we will link to code repositories and documentation where software used at OSU is deployable by others.

SKOS/FOAF Data Model: OSU Libraries chose a simplified data model to ensure easy maintenance of locally controlled names. The model uses SKOS properties (prefLabel, altLabel, and hiddenLabel) as its core name elements. FOAF is used for more detailed personal data (e.g. first name, last name, and date of birth) and for interoperability with non-library person data on the web. We present a full description and visual representation of the model and explain its relationship to MADS/RDF a more complex Linked Data authority system.

Linked Data Logistics: This work builds from a growing stack of Linked Data infrastructure at OSU Libraries. The presentation will offer some explanation what we had in place at the start of the project and the specific tools used. We will also address some challenges surrounding URI assignment and maintenance, which is a significant barrier for libraries publishing Linked Data, and data persistence.

DSpace Interoperability: Our system extends the built in DSpace authority control interfaces, implementing a generalized Linked Data authority service. The system uses the SPARQL protocol to access authorities from remote sources and translates the data for use by DSpace. We use the semantic web URIs as authority keys in the DSpace database. This service is designed to work with data conforming to the SKOS/FOAF data model, as well as the MADS/RDF authorities maintained by the Library of Congress.

Editor Application and Workflow: We developed a web application allowing staff to create, edit, and merge entities and their associated metadata. The application translates the SKOS and FOAF fields in the data model into a simple form, making it possible for staff to work with them without need for training on RDF or SKOS. The presentation will include a demonstration of this software.

At the time of submission, we are in the process of introducing staff to the editor application and establishing a workflow for maintaining authorities as a part of the DSpace review process. We will be able to report on the status of our workflow and provide estimates for the amount of staff time required.

Named Entity Matching: We undertook the work of controlling creator and contributor names after nearly eight years of successfully running our repository. This left us with a substantial cleanup project to disambiguate existing names. Briefly, we will address two approaches we have used to automate this cleanup. We first tried a naive approach, matching names based on simple heuristics. Though this effected a substantial improvement over the original data, we were eager to reduce missed and false matches. Second, we incorporated natural language processing tools for a more sophisticated approach. The focus, here, is on the requirements of repositories and lessons learned rather than a detailed treatment of the techniques.

Opportunities for Cross-Repository Authorities: Briefly, I will argue for a distributed authority system allowing repositories to share name records as Linked Data. This capability is implicit in the technologies used at OSU. Distributed authority would allow institutions to solve the name ambiguity problem more broadly, by focusing on controlling names specific or important to their materials and relying on others to do the same. This work may provide the foundation for such a solution.