GSoC 2011 - DSpace SKOS Authority Controls

Summary

Project	SKOS Authority Control Based on DSpace RDF Triplestore
Student	Yigang Zhou
Mentors	• Mark Diggory • Ryan Scherle
Technologies	SKOS RDF Triplestore, etc.
Proposal	Melange
Location for project	**(NOTE: Project was only partially completed)**

Project Description

@Mire has prototyped a Solr driven Authority Control capable of caching and mixing together authority sources so that they can utilized for super fast term completion and lookup. Solr is quite effective for quickly retrieving lists of values that a field should be restricted to. Likewise, when the original DSpace metadata is indexed into the Solr based authority control, the Submitter is also presented with an ad-hoc authority of existing values already contained within the repository. However, it is recognized that Authority Controls Sources are not just lists and have structural components as well. SKOS applies quite well to expressing the structured relationships between taxonomies and hierarchical vocabularies that are often relied on for Authority Control. Recent research in publishing Library of Congress Subject Authorities, Getty TGN Vocabularies etcetera confirms that SKOS is the predominant form to capture these resources for placement on the web.

Create a SKOS RDF Triplestore Authority Control for DSpace that utilizes SPARQL to provide a rich queriable local cache of Authority Control Sources that may be utilized in term completion and lookup in existing Authority Controls. Extend the Authority presentation to support more useful AC exploratory widgets using jquery and AJAX.

Refer to the following projects and resources for ideas:

- DSpace Sesame Triplestore
- DSpace Tupelo Storage Service
- The HIVE Project: https://www.nescent.org/sites/hive/Main_Page
- LoC SKOS Sources: http://id.loc.gov/

Project Significance

Years ago, neither the standards nor the software underlying institutional repositories anticipated performing Authority Control on widely disparate metadata from highly unreliable sources. Without it, though, both machines and humans are stymied in their efforts to access and aggregate information by author or the metadata field: 1) comparing plain text values can give false positive results e.g. when two different people have a name that is written the same. 2) it can also give false negative results when the same name is written different ways, e.g. "J. Smith" vs. "John Smith". Many organizations are awakening to the problems and possibilities of Authority Control. DSpace 1.6 provides a prototype of the new choice management and Authority control of Item ("DC") metadata values features, by introducing two new Metadata Fields: Authority Key and Confidence. However, it is recognized that Authority Controls Sources are not just lists and have structural components as well. SKOS, which is based on RDF, applies quite well to expressing the structured relationships (such as “border”, “narrower” and “related”) between taxonomies and hierarchical vocabularies that are often relied on for Authority Control. Recent research in publishing Library of Congress Subject Authorities, Getty TGN Vocabularies etcetera confirms that SKOS is the predominant form to capture these resources for placement on the web.

Approach

The semantic tripletore initiatives of my last GSoC project in 2010 pave the way for integration of DSpace and Fedora by providing a pluggable storage service tier for DSpace interacting with Mulgara, Sesame and other Triplestore implementations. This RDF triplestore solution for DSpace can be used to create a SKOS RDF triplestore Authority Control that utilizes SPARQL to provide a rich queriable local cache of Authority Control sources that may be utilized in term completion and lookup in existing Authority Controls. The Authority Control sources are abundant from the internet, e.g. Library of Congress Subject Authorities, provides all kinds of authorities and vocabularies, in SKOS (RDF/XML) format.

On the other hand, @Mire has prototyped a Solr driven Authority Control capable of caching and mixing together authority sources so that they can utilized for super fast term completion and lookup. Solr is quite effective for quickly retrieving lists of values that a field should be restricted to. Likewise, when the original DSpace metadata is indexed into the Solr based authority control, the Submitter is also presented with an ad-hoc authority of existing values already contained within the repository. In similar way, the project Hive (a SILS Metadata Research Center/NESCent collaboration) provides both SPARQL based semantic querying and Lucene indexing based keyword searching for Authority Control concepts in SKOS (RDF/XML) format. The success of project Hive has proved the feasibility of this GSoC 2011 project. Therefore I propose to use DSpace TupeloStorageService for SKOS Authority Control storage and SPARQL/Solr for fast term completion and lookup in existing Authority Controls.

Project Plan

I'm going to work on the design and development of SKOS RDF triplestore Authority Control storage and searching first (Week 1 – Week 7), followed by the development of the its presentation to support more useful Authority Control exploratory widgets (the rest of the weeks).

Week 1 (May 23 – May 29)

Tasks:

- design the API of SKOS RDF triplestore Authority Control storage and searching
- discuss the details of the implement of the API (i.e. feasibility, complexity and performance), and determine which approach to adopt.

Deliverable:

- Java API of SKOS RDF triplestore Authority Control storage and searching
- documentation of the API implementation architecture, approach and related technologies.

Week 2, 3 (May 30 – June 12)

Tasks:

- develop the storage part for SKOS RDF triplestore Authority Control.
- write unit tests

Deliverable:

- a component that can store SKOS RDF Authority Control sources in DSpace.
- test report of the component on different Authority Control sources

Week 4 (June 13 – June 19)

Tasks:

- develop the searching part for SKOS RDF triplestore Authority Control.
- write unit tests

Deliverable:

- a component that can query SKOS RDF Authority Control sources using SPARQL.
- test report of the component on different Authority Control sources

Week 5, 6 (June 20 --July 3)

Tasks:

- integrate Solr into SKOS RDF Authority Control storage component, to support fast indexing and searching for Authority Control concepts.
- write unit tests

Deliverable:

- Indexer and Searcher of Solr
- test report of the Indexer Searcher on different Authority Control sources

Week 7 (July 4 – July 10) - mid-term evaluation:

Tasks:

- clean up documentation
- prepare for the mid-term evaluations
- leave time room for manoeuvre, such as other project cooperation in "DSpace & Fedora Integration"

Deliverable:

- a component of SKOS RDF triplestore Authority Control for DSpace.
- detailed documentation of the implementation and related technologies.
- mid-term evaluation of this project

Week 8 (July 11 – July 17)

Tasks:

- study the UI documentation of DSpace, especially the original Authority Control UI.
- discuss with the mentors, and design the UI presentation of SKOS RDF Authority Control
- write UI demo tests

Deliverable:

- UI demo tests.
- documentation of the UI presentation design.

Week 9, 10 (July 18 – July 31)

Tasks:

- develop UI presentation of SKOS RDF Authority Control using JQuery and Ajax
- write unit tests

Deliverable:

- widgets or other UI components for SKOS RDF Authority Control searching.
- unit test report

Week 11 (August 1 – August 7)

Tasks:

- integration tests of the core component and the UI presentation of SKOS RDF Authority Control searching
- fix bugs according to the integration tests

Deliverable:

- integration test report

Week 12 (August 8 --August 14) - final evaluation:

Tasks:

- clean up documentation
- prepare for final evaluation to Google
- leave time room for manoeuvre, such as other project cooperation in “DSpace & Fedora Integration”

Deliverable:

- a component of SKOS RDF triplestore Authority Control for DSpace with presentation widgets supporting more useful AC exploratory.
- project full documentation

August 22 - Firm 'pencils down' date

Tasks

- submit final evaluations to Google

Deliverable:

- final evaluation of this project

August 26 - Final evaluation deadline