Authority Control of Metadata Values

Introduction

With DSpace you can describe digital objects such as text files, audio, video or data to facilitate easy retrieval and high quality search results. These descriptions are organized into metadata fields that each have a specific designation. For example: dc.title stores the title of an object, while dc.subject is reserved for subject keywords.

For many of these fields, including title and abstract, free text entry is the proper choice, as the values are likely to be unique. Other fields are likely to have values drawn from controlled sets. Such fields include unique names, subject keywords, document types and other classifications. For those kinds of fields the overall quality of the repository metadata increases if values with the same meaning are normalized across all items. Additional benefits can be gained if unique identifiers are associated as well in addition to canonical text values associated with a particular metadata field.

This page covers features included in the DSpace submission forms that allow repository managers to enforce the usage of normalized terms for those fields where this is required in their institutional use cases. DSpace offers simple and straightforward features, such as definitions of simple text values for dropdowns, as well as more elaborate integrations with external vocabularies such as the Library of Congress Naming Authority.

Simple choice management for DSpace submission forms

The DSpace Submission forms, defined in the submission-forms.xml file, allows the inclusion of value pairs that can be organized in lists in order to populate dropdowns or other multiple choice elements. If you explore the default submission-forms.xml file, you can see that a number of such value pair lists are already pre defined.

Example

```
<value-pairs value-pairs-name="common_identifiers" dc-term="identifier">
    <pair>
        <displayed-value>Gov't Doc #</displayed-value>
        <stored-value>govdoc</stored-value>
    </pair>
    <pair>
        <displayed-value>URI</displayed-value>
        <stored-value>uri</stored-value>
    </pair>
    <pair>
        <displayed-value>ISBN</displayed-value>
        <stored-value>isbn</stored-value>
    </pair>
</value-pairs>
```

It generates the following HTML, which results in the menu widget below.

```
<select name="identifier_qualifier_0">
    <option VALUE="govdoc">Gov't Doc #</option>
    <option VALUE="uri">URI</option>
    <option VALUE="isbn">ISBN</option>
</select>
```

A list of value pairs has following required attributes:

- **value-pairs-name** – Name by which an *input-type* refers to this list.
- **dc-term** – Dublin Core field for which this choice list is selecting a value.

Each *value-pairs* element contains a sequence of *pair* sub-elements, each of which in turn contains two elements:

- **displayed-value** – Name shown (on the web page) for the menu entry.

- **stored-value** – Value stored in the DC element when this entry is chosen. Unlike the HTML *select* tag, there is no way to indicate one of the entries should be the default, so the first entry is always the default choice.

Use simple choice management to add language tags to metadata fields

DSpace uses the simple choice management to provide a controlled list of language tags. Out of the box DSpace comes with a list of ISO language tags. You can add further language lists or use the provided one to let submitters tag languages of metadata fields. Take a look at the part of this documentation about the configuration of the [Submission User Interface](#).

Hierarchical Taxonomies and Controlled Vocabularies

The value pairs system works well for short and flat lists of choices. DSpace offers a second way of structuring and managing more complex, hierarchical controlled vocabularies. In contrast to the value pairs system, these controlled vocabularies are managed in separate XML files in the `[dspace]/config/controlled-vocabularies/` directory instead of being entered straight into submission-forms.xml

The taxonomies are described in XML according to this structure:

```
<node id="acmccs98" label="ACMCCS98">
    <isComposedBy>
        <node id="A." label="General Literature">
            <isComposedBy>
                <node id="A.0" label="GENERAL"/>
                <node id="A.1" label="INTRODUCTORY AND SURVEY"/>
                ...
            </isComposedBy>
        </node>
        ...
    </isComposedBy>
</node>
```

As you can see, each node element has an id and label attribute. It can contain the isComposedBy element, which in its turn, consists of a list of other nodes.

You are free to use any application you want to create your controlled vocabularies. A simple text editor should be enough for small projects. Bigger projects will require more complex tools. You may use Protegé to create your taxonomies, save them as OWL and then use a XML Stylesheet (XSLT) to transform your documents to the appropriate format. Future enhancements to this add-on should make it compatible with standard schemas such as OWL or RDF.

Default Hierarchical Controlled Vocabularies

By default, DSpace includes two out-of-the-box hierarchical controlled vocabularies in the `[dspace]/config/controlled-vocabularies/` directory.

- **nsi** - *nsi.xml* - The Norwegian Science Index (in the Norweigen language)
- **srsc** - *srsc.xml* - Swedish Research Subject Categories (in the English language, with notes in Swedish)

You may create your own hierarchical controlled vocabulary by using either of those as a model. All valid hierarchical vocabularies should align with the "controlledvocabulary.xsd" schedule available in that same directory.

Enabling / Disabling a Hierarchical Controlled Vocabulary

To enable a hierarchical controlled vocabulary, simply configure it's usage in one (or more) of your fields in your "submission-forms.xml" (as documented below).

To disable a hierarchical controlled vocabulary, simply remove it from all your fields in your "submission-forms.xml". You can also disable **all** controlled vocabularies by commenting out the "DSpaceControlledVocabulary" plugin in "authority.cfg":

authority.cfg

```
plugin.selfnamed.org.dspace.content.authority.ChoiceAuthority = \
  org.dspace.content.authority.DCInputAuthority, \
  org.dspace.content.authority.DSpaceControlledVocabulary
```

How to invoke a controlled vocabulary from submission-forms.xml

Vocabularies need to be associated with the correspondent DC metadata fields. Edit the file `[dspace]/config/submission-forms.xml` and place a "`vocabulary`" tag under the "`field`" element that you want to control. Set value of the "`vocabulary`" element to the name of the file that contains the vocabulary, leaving out the extension (the add-on will only load files with extension "*.xml"). For example:

```
<field>
    <dc-schema>dc</dc-schema>
    <dc-element>subject</dc-element>
    <dc-qualifier></dc-qualifier>
    <repeatable>true</repeatable>
    <label>Subject Keywords</label>
    <input-type>onebox</input-type>
    <hint>Enter appropriate subject keywords or phrases below.</hint>
    <required></required>
    <vocabulary>srsc</vocabulary>
</field>
```

The vocabulary element has an optional boolean attribute `closed` that can be used to force input only with the Javascript of controlled-vocabulary add-on. The default behaviour (i.e. without this attribute) is `closed="false"`. This allows the user to enter values as free text in addition to selecting them from the controlled vocabulary.

Authority Control: Enhancing DSpace metadata fields with Authority Keys

The aforementioned features only deal with text representations of controlled values. DSpace also offers support for adding authority keys and confidence values to a specific text value entered in a metadata field. The following terminology applies in the description of this area of DSpace functionality:

- **Authority** An *authority* is an external source of fixed values for a given domain, each unique value identified by a *key.* For example, the OCLC LC Name Authority Service, ORCID or VIAF.
- **Authority Record** The information associated with one of the values in an authority; may include alternate spellings and equivalent forms of the value, etc.
- **Authority Key** An opaque, hopefully persistent, identifier corresponding to exactly one record in the authority.

The fact that this functionality deals with **external** sources of authority makes it inherently different from the functionality for controlled vocabularies. Another difference is that the authority control is asserted *everywhere* metadata values are changed, including unattended/batch submission, SWORD package submission, and the administrative UI.

How it works

TODO

Original source:

Authority Control of Metadata Values original development proposal for DSpace 1.6