Controlled Vocabularies and Authority Control in DSpace 7

> ⊘ **This page is outdated / obsolete**
>
> This page represented a temporary discussion area around issues found in how these features were implemented in an early version of DSpace 7 (pre-7.0).
>
> A refactoring of these features has been designed/moved forward into this REST Contract PR: https://github.com/DSpace/Rest7Contract/pull /128. This is planned to be implemented by DSpace 7.0 beta 3
>
> Therefore, this page is now OBSOLETE and the REST CONTRACT should be considered the latest version.

> ⚠ Draft notes (not complete or finalized). Based loosely on conversation in https://github.com/DSpace/DSpace/pull/2743

How this worked in DSpace 6

Three concepts:

- **Authority control values** *tend to be external.* It was originally designed to search external sources like Library of Congress or SHERPA Romeo
- **Controlled vocabularies** *tend to be internal.* They exist in a configuration subfolder with a specific XML schema: https://github.com/DSpace /DSpace/blob/dspace-6_x/dspace/config/controlled-vocabularies/
    - Samples include SRSC: https://github.com/DSpace/DSpace/blob/dspace-6_x/dspace/config/controlled-vocabularies/srsc.xml
- **Value-Pairs** are a "controlled vocabulary" like concept that exist only in the submission configuration (input-forms.xml): https://github.com/DSpace /DSpace/blob/dspace-6_x/dspace/config/input-forms.xml#L236
    - Samples include "common_identifiers" (URI, ISSN, ISBN, etc), "common_types" (Article, Book, Image, Video), "common_iso_languages" (en, es, de, fr, etc)

Default settings:

1. Value-pairs were *always enabled* and act like a simply form of controlled vocabulary. They were only used to populate dropdowns in submission.
2. Additional controlled vocabularies were *disabled by default*. Though they could be enabled for JSPUI only: https://github.com/DSpace/DSpace /blob/dspace-6_x/dspace/config/dspace.cfg#L1809
3. Authority control was also *disabled by default,* though a variety of plugins were available out-of-the box for Library of Congress, SHERPA Romeo and ORCID: https://github.com/DSpace/DSpace/blob/dspace-6_x/dspace/config/dspace.cfg#L1432-L1437
    a. It was also possible to optionally configure Authority Control to **use value-pairs and/or controlled vocabularies.** See https://github.com /DSpace/DSpace/blob/dspace-6_x/dspace/config/dspace.cfg#L1462-L146
        i. This is disabled by default. When enabled, both value-pairs and controlled vocabularies could be used as authority control on a specific field(s).

How this is working in DSpace 7.0 beta2

1. *(Same/Similar behavior)* Value-pairs are still always enabled in the new submission-forms.xml: https://github.com/DSpace/DSpace/blob/master /dspace/config/submission-forms.xml#L1130
2. *(Different behavior)* Additional controlled vocabularies are **enabled by default,** as they are configured to be used in Authority Control: https://githu b.com/DSpace/DSpace/blob/master/dspace/config/dspace.cfg#L1483
3. *(Different behavior)* Authority control is **enabled by default**, and defaults it using **both value-pairs** and **controlled vocabularies:** https://github. com/DSpace/DSpace/blob/master/dspace/config/dspace.cfg#L1481-L1483
4. *(Different behavior)* The Submission configuration/UI now **only uses Authority Control for dropdowns**, as both value-pairs and controlled vocabularies are authority control enabled. See early notes at Configuration changes in the submission process
    a. *(Same/Similar behavior)* Controlled Vocabularies can be associated with a field via the <vocabulary> tag in submission-forms.xml: https:// github.com/DSpace/DSpace/blob/master/dspace/config/submission-forms.xml#L202
    b. *(Same/Similar behavior)* Value-pairs can be associated with a field via the <input-type value-pairs-name="name">dropdown</input-type> tag. For example: https://github.com/DSpace/DSpace/blob/master/dspace/config/submission-forms.xml#L180

These behaviors primarily changed in https://github.com/DSpace/DSpace/pull/1852/

Side effects of new behavior in 7.0 beta2

1. Because all value-pairs and controlled vocabulary are authority controlled by default, all metadata fields generated from either now have "authority IDs". This makes them look/act different from DSpace 6 metadata.
    a. For value-pairs, the authority ID is the stored value. So, e.g. "Book" has an authority ID of "Book". See examples in REST Contract: https ://github.com/DSpace/Rest7Contract/blob/master/authorities.md#authority-entry-values
    b. Other values in the same field don't have an authority ID, e.g. values from an upgraded DSpace 6 or values from another collection which uses the same field as plain text, or values submitted through SWORD, …
    c. *This will have a side effect for some other tools.*
        i. For example, Batch Metadata Editing separates authorities from values via a "::". So, a dc.type value of "Book" will now appear as "Book::Book" in a CSV exported spreadsheet (whereas it used to appear just as "Book")
        ii. Discovery has always treated authority values separate in both search and browse. Values with an authority ID are displayed separate from values without an authority ID

           iii. OAI-PMH uses XML following the XOAI schema. This XML is run through XSLT to generate the XML to provide for different metadata formats in OAI-PMH. Values coming from authority control are provided different in XOAI than "normal" metadata values. Either a lot or institutions have to change their XSLTs or we do change how XOAI XML deals with authority control.

2. The Authority ID doesn't add any additional information to the field. In fact, it's **not a unique Identifier.** As a basic example, both the value-pairs for "common_identifiers" and "common_types" include an "Other" option. So a value linking to Authority ID "Other" cannot be mapped back to its source without finding the "submission-form.xml" which was used to create it (and that configuration may have been modified since then).

     a. This behavior was identical in DSpace 6, if you used value-pairs as Authority Control. But, as noted above, this feature was disabled by default.