Controlled Vocabularies (CV) Decision Tree

September 2020; updated April 2021 to add citation link and clarify Terms List examples
Dawn Childress, Gretchen Gueguen, Julie Hardesty, Rebecca Pattillo, Erik Radio, Lynette Rayle

Introduction

Discovery and retrieval in various digital platforms and repositories can be greatly enhanced by the use of controlled vocabularies (CV). Controlled vocabularies are sets of terms that can be used to ensure greater consistency among metadata records, as opposed to uncurated metadata values. This consistency allows a user to collocate, filter, and compare resources and facilitates their ability to find ones that they need. While commonly used for subjects, genres, and other categorical fields, CVs can also be used for rights statements, names, places, and other metadata fields for which more consistency can enable better system functionings and discovery.

While there are many well-established CVs in use (e.g. Library of Congress Subject Headings; Getty Art and Architecture Thesaurus), it is important to remember that CVs are works in progress, being continuously updated to better reflect the areas and disciplines they help describe. Similarly, CVs can also be a tool for marginalized communities who historically have not had the power to describe or shape their representation in the world. As such, the use of a particular CV and its impact on the description of resources in a given collection warrant evaluation to determine if the CV is a good match for a set of resources. See the accompanying Controlled Vocabularies and Criteria for CV options and consider adding any CVs you know about that are not listed.

Scope

This decision tree is meant to provide guidance for evaluating, selecting, and using controlled vocabularies for descriptive metadata fields. Using a tiered approach, the document explores three scenarios for using CVs. First, it discusses how to evaluate a CV already built into a given platform. Second, it looks at replacing a CV in a platform with another existing one. Third, it addresses when it is necessary to customize, modify, or build a local CV due to gaps or deficiencies in existing CVs.

Audience

While this document is written for Samvera specific implementations (e.g. Hyrax/Hyku), many of its considerations may be applicable to other repository platforms. Similarly, this document will be useful for Developers, Metadata/Cataloging Librarians, Digital Collection or Repository Managers, and others who create and maintain digital repository resources. The use cases provided in this document are generalized examples of fields that commonly use controlled vocabularies to enable better indexing, search, and discovery. While this document gives tips and suggestions for the evaluation, selection, and use of CVs, it is not its aim to be prescriptive in these criteria, nor provide in-depth instruction for the development of local CV. To this end, readers are encouraged to examine the *Further Reading* section for additional literature on the theory and implementation of CVs more broadly.

Decision Tree Levels for Controlled Vocabularies

Level 1: Working with your Content Management System's (CMS) existing controlled vocabularies

When migrating to a new system or setting up a new digital collection or institutional repository, understanding and evaluating the existing controlled vocabularies that are integrated into the software or CMS is important. Some questions to consider include:

- Which, if any, CVs are already utilized in the descriptive metadata?
- Which CVs are available in the software?
- Do the built-in CVs comply with an established descriptive metadata standard?

Tips for Evaluating CVs

Evaluate the existing CVs in relation to the collection by considering:

- The *format* of materials in the collection (i.e. artworks, manuscripts, photographs, audio/video, research data, newspapers, etc.)
- The *subject* of the materials in the collection (i.e. art, educational, local history, communities of people, medical, scientific, etc.)

With the above in mind, do the existing CVs provide the most useful terms for search and discovery, indexing, and/or sorting? If not, consider connecting to an external CV, such as LCNAF or Getty's AAT, or creating your own locally defined Term List for use in your content management system. (see **Levels 2 and 3 below).**

What is a locally defined Term List?

This document distinguishes between a **CV** and a locally defined **Term List.** A CV, for the purposes of this document, is an authoritative resource created by a community for shared use to provide consistent and common metadata terms. Library of Congress Subject Headings is an example of a CV as is the North Carolina Council on Developmental Disabilities' Glossary of Disability Terms. A Term List, while technically a type of CV, is used here to denote a list of terms defined and used locally and likely not published or available online for others to use. An internal set of terms for use in an application, such as a list of resource types, or a list of department names for an organization are examples of a Term List. Collections can use a combination of CVs and Term Lists.

Example in Hyrax/Hyku: Existing CVs and Term Lists

Hyrax/Hyku comes with a few predefined sources for metadata values. For more details, see Appendix 1: Out-of-the-Box Hyrax/Hyku Fields that Use CVs:

- The "Location" field uses Geonames, a community CV for geographic and place names;
- The "Resource Type" field uses a predefined Term List;
- The "Rights Statement" field has a pre-populated Term List stored locally within the application and based on RightsStatements.org;
- The "License" field uses a pre-populated Term List of Creative Commons values, also stored locally within the application

Level 2: Replacing controlled vocabularies

If a review of the CMS's controlled vocabularies shows that these will not be suitable, explore options for replacing the existing controlled vocabularies with alternatives. Considerations here include:

- Are there particular formats required for controlled vocabularies by the software?
- Are instructions available and useful for replacing one controlled vocabulary with another in the platform?
- What other controlled vocabulary options exist and are they more suitable?

When looking for other controlled vocabulary options, consider who is creating a controlled vocabulary, what that controlled vocabulary represents, and how it is being used elsewhere.

- Consider if there is an organization that represents a community / industry / profession relevant to the collection(s). Do they publish or endorse a standard?
- Consider cultural heritage institutions that might have similar collections described online. What do they use for their descriptions?
- Reach out to other professionals who may have experience. There are many groups that exist for metadata support including:
    - Digital Library Federation (DLF) Metadata Support Group on Slack
    - DLF Metadata Assessment Working Group
    - Samvera Metadata Interest Group
    - SAA Metadata and Digital Objects Section

Tips for Evaluating Replacement CVs

Evaluate CV replacement options in relation to your collection by considering:

- Who created the CV? Are they a group that is established and relevant to the domain? Do they represent the community / industry / concern served by the CV?
- Does anyone maintain the standard? When was the last time it was updated?
- Are there stable identifiers or URIs available?
- What domain does the CV represent, and are the vocabulary terms relevant to it? Is a desired term included in the CV, but generally outside of the domain of it?
    - For example, a location term within a non-location-based CV could be better represented by a CV with the sole purpose of providing locations.
    - As another example, does the CV include terms about a community or specific group but is not created or maintained by that community or group? Does that community or group provide a separate vocabulary?
- In reviewing the vocabulary, does it seem complete and well constructed?
- In what format does the CV exist? Is it available as a linked data endpoint or other searchable resource? Does it need to be formatted or transformed to be able to use?
    - For example, is it already possible to use this in Hyrax/Hyku via Questioning Authority - QA?

Appendix 2: Types of Fields that Commonly Use CVs offers examples of types of fields that might use CVs and some CV examples that can be applied.

Example in Hyrax/Hyku: Replacing CVs

As an example, consider a scenario where the collections being managed in Hyrax/Hyku includes LGBTQ+ collections. Homosaurus.org is an international LGBTQ Linked Data vocabulary and would be helpful to use for applying Subject field terms. This CV works as a replacement CV since it is updated, maintained, and used by the Digital Transgender Archive, an active online archive, to be "a robust and cutting-edge vocabulary of LGBTQ-specific terminology that enhances the discoverability of LGBTQ resources." The CV is regularly updated and is available online as Linked Data.

For information on how to replace a CV for a particular field, see the Samvera Knowledge Base on Modifying the Edit Form, which includes information on using the Questioning Authority gem (QA) to incorporate Linked Data CVs: https://samvera.github.io/customize-metadata-edit-form.html. While Homosaurus.org is not currently available as part of the QA gem, it is available online as Linked Data and can be used with QA.

Level 3: Customizing, modifying, and creating controlled vocabularies for specific use cases

In some situations an appropriate CV is available, but doesn't quite fit the system's requirements or may require a few modifications for the purpose of the project. In these cases, options include modifying the existing CV or creating an entirely new one.

Modifying a CV

When customizing or modifying a controlled vocabulary for local use, consider the following:

- Is the CV available for repurposing? Is permission or licensing required by the CV maintainer? This can apply to simply changing the CV format to use with particular software, or may be needed to modify the terms themselves.
- What are the steps involved to make any changes?
- What are the steps involved to maintain this customized/modified CV locally? How will the customized/modified CV be maintained long term?

Creating a new Term List

When no existing CV exists that meets project requirements, a local CV or Term List can be created. When a larger domain of knowledge or class of materials is not covered by an existing CV, the harder work of developing a full taxonomy that can be shared outside the project may be in order. This is no small task and is usually undertaken by professional groups.

In most cases, however, a local Term List will be sufficient for project purposes. For example, an institutional repository may want to capture university specific terms such as school name, department name, unit name, degree conferred, etc. Since this is a relatively small set of terms, and is only applicable locally, building a Term List and incorporating it into your system is the best option.

Tips for Creating Local Term Lists

- See Further Reading available at the end of this document for resources to consult on developing a new CV or Term List
- Think about future migration of the list. If this Term List is created and maintained locally, how will it be documented and managed? Future updates to CMS software might interfere with using a Term List so providing good documentation will help with any potential migration issues.
- Document the purpose of the Term List and how it should be used to prevent confusion in future projects.
- Sharing resources outside of the system, such as with a state, regional, or national hub, may require stable identifiers over time. Consider creating stable identifiers like URIs that will not change over time and can be easily passed along with other metadata.

Example in Hyrax/Hyku: Modifying or creating a Term List

Creating a Term List: As an example, consider the scenario mentioned above of creating a list for an Institutional Repository of schools, colleges, and departments within a university. The first step might be to identify the official list of departments and programs from the university, and then determine the level of granularity of the list (Is the department name enough? Do specific programs need to be articulated?). Once a list has been developed, follow the instructions at https://samvera.github.io/customize-metadata-controlled-vocabulary.html or https://samvera.github.io/customize-metadata-model.html as appropriate. Be sure to create documentation on how the list was created as well as how to use it. These may be added to existing metadata guidelines or created on their own.

Modifying a Term List: Modifying an existing term list in Hyrax/Hyku may require working with your system developers. Appendix 1 provides examples of fields that come with CVs in Hyrax/Hyku that can be modified. For example, you can remove a Rights Statement option like “Copyright Not Evaluated” by removing the id:, term:, and active: properties from the rights_statement.yml file. However, customizing controlled vocabularies through Questioning Authority (QA) is possible. See Appendix 3: CVs available with Questioning Authority (QA) for CVs currently available with Questioning Authority.

Further Reading

Overview

ANSI/NISO. (2010). Z39.19-2005 (R2010) Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. https://www.niso.org/publications/ansiniso-z3919-2005-r2010

Harpring, P. (2010). What Are Controlled Vocabularies? *Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works. Los Angeles, CA: Getty Publications*. https://www.getty.edu/research/publications/electronic_publications/intro_controlled_vocab/what.html

Controlled Vocabulary Design

Mai, J. E. (2008). Actors, domains, and constraints in the design and construction of controlled vocabularies. *KO KNOWLEDGE ORGANIZATION*, *35*(1), 16-29. https://www.nomos-elibrary.de/10.5771/0943-7444-2008-1-16/actors-domains-and-constraints-in-the-design-and-construction-of-controlled-vocabularies-volume-35-2008-issue-1

Svenonius, E. (1989). Design of controlled vocabularies. *Encyclopedia of library and information science*, *45*(suppl 10), 82-109. https://www.semanticscholar.org/paper/Design-of-Controlled-Vocabularies-Svenonius/4ab13a4e1a199ce5b05e5530dbb61946eae09ff1

Creating Controlled Vocabularies

5 Star Open Data. (2012). https://5stardata.info/en/

BC First Nations Subject Headings. http://branchxwi7xwa.sites.olt.ubc.ca/files/2011/09/bcfn.pdf

Indigenous Subject Headings Modifications https://www.youtube.com/watch?reload=9&v=a_-G9MNBL6E&feature=youtu.be

Kaltman, E., Wardrip-fruin, N., Mastroni, M., Lowood, H., De groat, G., Edwards, G., ... & Caldwell, C. (2016). Implementing controlled vocabularies for computer game platforms and media formats in SKOS. *Journal of Library Metadata*, 16(1), 1-22. https://www.tandfonline.com/doi/full/10.1080/19386389.2016.1167494

Littletree, Sandra, and Cheryl A. Metoyer. 2015. “Knowledge Organization from an Indigenous Perspective: The Mashantucket Pequot Thesaurus of American Indian Terminology Project.” *Cataloging & Classification Quarterly* 53 (5–6): 640–57. https://doi.org/10.1080/01639374.2015.1010113.

Riley, Jenn. (2010). Seeing Standards: A Visualization of the Metadata Universe. http://jennriley.com/metadatamap

W3C. (2008). Best Practice Recipes for Publishing RDF Vocabularies. https://www.w3.org/TR/swbp-vocab-pub/

Welhouse, Z., Lee, J. H., & Bancroft, J. (2015). “What Am I Fighting For?”: Creating a Controlled Vocabulary for Video Game Plot Metadata. *Cataloging & Classification Quarterly, 53(2), 157-189.* https://www.tandfonline.com/doi/full/10.1080/01639374.2014.963776

Appendices

Namespace Key

Prefix	Namespace
dc / dce	http://purl.org/dc/elements/1.1/
dct / dcterms	http://purl.org/dc/terms/
edm	http://www.europeana.eu/schemas/edm/
foaf	http://xmlns.com/foaf/0.1/
relators	http://id.loc.gov/vocabulary/relators

Appendix 1: Out-of-the-Box Hyrax/Hyku Fields that Use CVs

The following fields have CVs in place when Hyrax/Hyku is installed and started.

Field name	Property	Literal/URI	Hyrax technical approach
License	dct:rights	URI	QA via Hyrax yml file
Location	foaf:based_near	URI	QA access to Geonames
Resource type	dct:type	Literal	QA via Hyrax yml file
Rights statement	edm:rights	URI	QA via Hyrax yml file

Appendix 2: Types of Fields that Commonly Use CVs

General Concept	Vocabulary source examples	MODS/RDF predicate
genre	Getty AAT	edm:hasType
	LCGFT	
language	ISO639-2 Codes	dcterms:language
license	creativecommons.org	edm:rights
location	Geonames	dce:coverage
resource type	DCMI Type Vocabulary	dcterms:type
rights	rightsstatement.org	edm:rights
subject	LCSH	dce:subject
	OCLC FAST	
	MeSH	
	Homosaurus	
name	LCNAF	relators:[term]

	ULAN	dce:creator
	VIAF	dce:contributor
	ORCID	
department/unit	local	
uniform title	LCNAF	dce:title

Appendix 3: CVs available with Questioning Authority (QA)

Direct access to authority providers via non-linked data APIs...

Authorities	**Subauthorities**	**comments**
Discogs	master, release, all	
GeoNames	N/A	
Getty	aat, tgn, ulan	
Library of Congress (LOC)	many, see /lib/qa/authorities/loc_subauthority.rb	
Medical Subject Heading (MeSH)	N/A	
OCLC FAST	all, personal, corporate, event, uniform, topical, geographic, form_genre	

Direct access to authority providers via linked data APIs...

Authorities	**Subauthorities**	**comments**
Agrovoc	N/A	
DBPedia	N/A	
Geonames	N/A	
Library of Congress (LOC)	names, subjects, genre, classification, child_subject, demographic, music_performance	single term dereferencing only
NALT (agricultural thesaurus)	N/A	
OCLC FAST	person, organization, topic, event_name, geocoordinates, uniform_title, period, form, alt_lc	

Downloads of linked data available for caching...

Once a local cache exists that supports a REST search API returning linked data, it is accessible through QA. The following have been cached as part of the LD4P grant and are in use now through the grants QA Lookup service.

For exploration only: https://lookup.ld4l.org/authority_list

Authorities	**Subauthorities**	**comments**
Agrovoc	N/A	
CERL	person, imprint, corporate	
DBPedia	N/A	
Geonames	area, place, area_and_place, water, park, road, spot, terrain, undersea, vegetation	
Getty AAT	many, see LD4P config	
Getty TGN	N/A	
Getty ULAN	person, organization	
Ligatus	N/A	
LOC Demographics	N/A	
LOC Genres	active, deprecated	
LOC Names (no RWO)	geographic, conference	Used for LCNAF entities that do not have real world objects (RWO)
LOC Names (with RWO)	person, organization, family	Used for LCNAF entities that do have real world objects (RWO)
LOC Performance	N/A	
LOC Subjects	N/A	
Medical Subject Heading (MeSH)	subject, publication_type	
NALT (agricultural thesaurus)	N/A	
OCLC FAST	concept, event, person, organization, place, intangible, work	
RDA Registry	many, see LD4P config	The subauthorities are the actual RDA Registry vocabularies (e.g. aspect_ratio, collective_title, etc.). This does not support search across subauthorities.
ShareVDE (one per	work, superwork, instance	LD4P has cached 20 institutions data

institution)		from Share VDE. Each institution is a separate authority.