Using Normalized Subject Headings from CDI

Introduction

CDI Subject terms can come from many different sources and providers, and in various styles and formats. Some CDI sources use a subject vocabulary, while in other cases, the CDI subjects are author provided keywords, or they are subjects that come from various sources (such as from ‘’aggregator’’ databases).

To provide more consistent, deduplicated, and normalized subjects across the CDI index, the existing Subject field has been divided into the following fields:

- A (Normalized) Subject field that includes only subjects that match the CDI controlled vocabulary. They are deduplicated and normalized.
- A Keyword field that includes all subjects that could not be mapped against the controlled vocabulary. These subjects will be considered keywords.

The CDI controlled vocabulary uses the preferred terms and variant terms from the following sources:

- Library of Congress Subject Headings (LCSH) – information at https://id.loc.gov/authorities/subjects.html
- MeSH – information at https://www.ncbi.nlm.nih.gov/mesh/
- A small subset of Proquest thesaurus terms

For more information about the CDI controlled vocabulary, see Rules for Adding Subjects to the Normalized Subject Field.

When enabled, this functionality provides the following:

- Only the normalized subjects appear for the Subject/Topic facet and the Subject display field.
- When performing a search in CDI scopes, both the Subject and Keyword fields are searched to ensure that nothing is lost (since not all subjects can be mapped to the controlled vocabulary).

Note

- The first iteration of the CDI controlled vocabulary uses LCSH, MESH, and the Proquest thesaurus, which are based on the English language. We plan to extend the vocabulary over time, possibly from more thesauri and multilingual resources, which is subject to further analysis.
- Other future plans include identifying and removing disciplines from the Subject and Keyword field and then adding them to a separate Discipline field.
- At this time, we will not add subjects/keywords to a record unless the record already contains subjects/keywords from one of its original data sources. The source for populating the normalized Subject field remains the content provider’s data.

Using the New Subject and Keyword Fields

Search and Ranking

The subject facet is based on the normalized Subject field and does not include keywords. The search engine functions as before and gives the normalized Subject field special weight in the dynamic rank. The Keyword field is given less weight than subjects but is weighted above the abstract field (for example). In general, the ranking works the same as it did previously.

Display

When switching to the new field, the Subject facet is automatically populated from the normalized subjects, and the Keyword facet is automatically populated with the subjects that could not be normalized. Both the Subject and Keyword facets are configurable for display. For more information, see the following settings in Summon Admin Console:

- Subject Terms > Use Normalized CDI Subjects
- Quick Look > Display Keywords

The Normalized Subject Field

The normalized Subject field uses CDI's subject controlled vocabulary, which consists of subjects and their alternate terms. The subjects are compiled from a combination of the following sources:

- LCSH (Library of Congress Subject Headings) – The primary source for subjects and their alternate terms.
- MeSH (Medical Subject Headings) – The mapping file compiled by Northwestern University Libraries to avoid duplicates with LCSH. Mapping information can be found at: https://galter.northwestern.edu/about-us/northwestern-university-libraries-lcsh-mesh-mapping-project
- ProQuest Thesauri for a select set of additional terms that are very closely matched to LCSH.

In case there are duplicates, the LCSH terms take precedence.

Rules for Adding Subjects to the Normalized Subject Field

General

The subjects in the source record are normalized by converting uppercase letters to lowercase, removing extra spaces and punctuation, and then mapping the subjects against the CDI controlled vocabulary. If a subject can be mapped, it will be added to the normalized Subject field.

There are some exceptions that apply to all vocabularies used, for example—we are not adding very general terms such as **Style**, **Intention**, or content types (such as **Electronic books**). Variants are added as alternate terms for the subject, unless they are programmatically determined to be mappings to narrower subjects, mappings to broader subjects, or ambiguous subjects.

LCSH

LCSH is the primary source for the CDI controlled vocabulary subjects and their alternate terms.

For example, the LCSH subject "Driving of horse-driven vehicles" (sh85039614) has the following variants:

- Driving
- Driving, Horse-drawn vehicle
- Horse-drawn vehicle driving

Because the "Driving, Horse-drawn vehicle" variant maps to a narrower subject, it is excluded. The other variants are listed as alternate terms for this subject.

For LCSH complex subjects, CDI adds all components to the new normalized Subject field that exist as LCSH subjects. LCSH Complex subjects contain multiple components typically connected with a double dash, for example—"Japanese American -- Alcohol Use" (sh2008009026). In this example, both "Japanese Americans" (sh85069603) and "Alcohol Use" (sh99002331) are LCSH subjects, so they are both included in the new normalized Subject field.

This is different from the following example: "Japanese Americans--Forced removal and internment, 1942-1945" (sh85069606). Only its first component "Japanese Americans" (sh85069603) exists as a LCSH subject (and is included in the new normalized Subject field), but because the second component "Forced removal and internment, 1942-1945" does not exist as a LCSH subject, it is not included in the new normalized Subject field, but instead it is included with the Keyword field.

MESH and the Proquest Thesaurus

While LCSH serves as the primary source for the CDI normalized subjects, MeSH headings are used to add additional subjects and alternate terms that do not exist in LCSH. CDI uses the mapping between LCSH and MeSH from Northwestern University Libraries, which can be found at https://galter.northwestern.edu/about-us/northwestern-university-libraries-lcsh-mesh-mapping-project

MeSH subjects and their entry terms are added as alternate terms for the corresponding LCSH subject if there is a mapping entry in the Northwestern University Libraries file. If they cannot be mapped to an LCSH subject, MeSH terms are added as new subjects to the CDI list of subjects. Their entry terms are added as the alternate terms for the subject unless they are programmatically determined to be mappings to narrower subjects, mappings to broader subjects, or ambiguous subjects. For example, "Calcimycin" (D000001) is added as a subject, and "Antibiotic A23187" is added as its alternate term.

In addition, we use the ProQuest Thesaurus to include a small set of additional alternate terms that closely match the LCSH normalized subjects.