Subject Access to Digital Collections in CONTENTdm:
Background and Rationale for UO Decisions

9/27/04 cgh; rev. 12/28/05

Best practice for subject terms used in descriptive metadata for digital materials is to base the terms on a controlled vocabulary list. The choice of a particular controlled vocabulary may depend as much on the software being used and its limitations or features as on the target audience or the nature of the materials in the collection.

Software considerations

Within the U.S. library community, there is widespread use and acceptance of the Library of Congress Subject Headings (LCSH) and the Library of Congress Name Authority File (LCNAF) for controlled vocabulary for subject terms. However, the use of LCSH within library catalogs relies upon the MARC format which parses different types of subject access very finely (personal names, corporate names, geographic names, topical subjects, and assorted subdivisions) and the decades-long development of online catalog functionality to make good use of those fine distinctions. The library community also has detailed content standards for the type and structure of data to be input into different MARC fields. Online catalog features developed over three decades include the ability to conduct searches limited to certain types of subject information and, more importantly, the ability to provide cross references and redirection of searches from a user's search term to the official, established term. Even with such a highly developed infrastructure and tradition, library users are often confused by the standards and resort to keyword searching rather than subject searches.

The software being used for describing and making digital collections available is not as well-developed as library catalogs. Additionally, it is designed to support collections being built by organizations other than libraries. Making use of controlled vocabularies that were developed for use within one particular tradition, i.e. the MARC format and online library catalogs, is challenging without the supporting system functionality and its utility is questionable.

CONTENTdm software, used to build many of the digital library collections at the UO, does not support the MARC format. Instead, fields are mapped to Dublin Core (DC). Subject information is mapped to DC Subject, whether the field contains a personal name, corporate name, geographic name, time period, or topical term. Within CONTENTdm, there is only the most basic type of cross referencing available from an unused term to the official, used term, without any ability to show relationships between terms. In addition, the concept of subfield coding which functionally supports pre-coordinated strings of terms within the MARC format is not available within CONTENTdm. Searching is either keyword (in fields where a controlled vocabulary has not been used) or by phrase (in fields where a controlled vocabulary has been activated).

Due to the reduced functionality of these systems compared to an online library catalog, and the necessity of mapping to a metadata schema with fewer options for parsing distinct classes of subject data, the University of Oregon Libraries have made the following decisions regarding subject access:

General policies

Division of the world

In collections where different types of subject terms are entered in separate fields (personal names, corporate names, topical terms, etc.) we usually follow the Library of Congress' "division of the world" as outlined in Instruction Sheet H405 of the *Subject Cataloging Manual: Subject Headings (SCM).*

Cross references

We do not automatically include every 4xx that appears in an LC name or subject authority record or that appears as a cross reference in any other controlled list of terms. Nor do we necessarily follow SCM or AACR2 guidelines for the form of cross references. Instead, we are guided by the utility of a particular cross reference to our user communities and by the functionality of the software package(s) by which we are delivering digital content.

Topical terms

Topical terms are mapped to DC Subject.

Topical terms will be drawn from LCSH, AAT, TGM, or other source vocabularies. The source of subject terms for each collection will be documented at the collection level.

LCSH

For all collections begun since January 2004, we use a controlled LCSH subject list created from subject authority records currently in use in our online catalog and loaded into our CONTENTdm collections. This vocabulary includes only headings appearing in 150 fields of authority records. *All terms that appear as 150 fields in authority records are eligible for LC Subject fields in digital collections.*

- Subfields $x and $v will be retained but will be translated into two dashes to simulate the way these headings appear in an online catalog.

- Any authority records with a 150 $z will be stripped from the list, since geographic names are being handled in a separate field.

- The LC subject vocabulary within our digital collections includes selected cross references appearing in 4xx fields of LC subject authority records.

Because LCSH is a dynamic list, and because we are using only the subset of LCSH represented in our online catalog, we will regenerate the list periodically, load the new list into CONTENTdm, and perform any necessary clean-up of the LC subject terms applied to individual items in our digital collections. Our database maintenance staff who do this work in the online catalog will carry out this work within the CONTENTdm system.

TGM

CONTENTdm supplies the TGM vocabulary in its standard toolkit. TGM has particular value for image-based collections. Because there is a lot of overlap between LCSH and TGM, the need to supply both LCSH and TGM terms will be decided for each collection individually. When it has been determined that supplying TGM terms adds value to a digital collection, it will be applied as it is provided in the CONTENTdm software. No expansion of TGM base terms will be undertaken to create a pre-coordinated string. At this time, we do not plan to perform clean-up of previously valid but now obsolete TGM terms in our digital collections.

AAT, etc.

Decisions regarding the use of other source vocabularies will be made on a collection-by-collection basis.

Local vocabularies

Because of the specialized and highly focused nature of some of our digital collections, many of the existing controlled vocabulary lists do not include all terms that are needed to describe the materials in the collections. When local terms are used, we will attempt to document their source.

Personal, corporate, and conference names

When used to indicate what the digital object is of or about, personal, corporate, and conference names are mapped to DC Subject. Catalogers should consult the "Personal, Corporate, or Conference Names in Digital Collections" document (revised December 2005) for guidance on the form of names to be used in digital collections.

For earlier collections, different standards were sometimes followed. If time permits, names in earlier collections may be revised to conform to current practice.

For some specialized collections, such as collections where the primary target audience is composed of faculty and students from a particular discipline, names may be established using different guidelines. In those cases, the deviation from standard practice will be documented.

Place names

Although the DC Metadata Element Set indicates that geographic characteristics of digital objects should be mapped to Coverage.spatial, we have decided that we will map place names that describe what the object is of or about to DC Subject. If needed, we will duplicate geographic information in a separate field that is mapped to DC Coverage.spatial.

Place names are always entered in a separate field, rather than being added as part of a term in a topical subject field (no $z subfield equivalents).

Place names are always entered in direct order, from smallest jurisdiction to the larger jurisdiction.

Forms of place names are researched in LCNAF, GNIS, or the Columbia Gazetteer. If found in LCNAF, the form of name in LCNAF will be used. Names not found in LCNAF will be entered following AACR2 guidelines in Chapter 23, ***with the following exceptions:***

- No abbreviations will be used. Instead, the name of small and large jurisdictions will always be spelled out. For instance, we will use the place name **Pendleton, Oregon** rather than **Pendleton (Or.)**. This decision has been made because we believe that it better serves the potential users of these collections.
- Cross references will be entered as needed to the controlled list of names.
- In cases where an intervening jurisdiction is considered an important search term, it will be entered as a separate geographic name. For instance, Telephone Ridge is on the Umatilla Indian Reservation in Oregon. We will use the term **Telephone Ridge, Oregon** as well as providing an additional place name for **Umatilla Indian Reservation, Oregon.**

Time periods

Although the DC Metadata Element Set indicates that chronological characteristics of digital objects should be mapped to Coverage.temporal, we have decided that we will map time periods that describe what the object is of or about to DC Subject. If needed, we will duplicate chronological information in a separate field that is mapped to DC Coverage.temporal.

Time periods that describe what the object is of or about may be taken from those established by the Library of Congress or they may simply represent the date that appears on the original object.

Place names with subdivisions

Topical and chronological subdivisions may be used with the place name. Occasionally a digital object depicts or relates information about a place in a particular time period. Such chronological designations will be taken from those used in the LCNAF. This will be mapped to DC Subject, even if it is put into a separate field by itself rather than being linked to a place name.

- Subfields $x and $y will be retained but will be translated into two dashes to simulate the way these headings appear in an online catalog. For instance, if an image depicts a political scene from China in 1913, an appropriate subdivision string may be included as part of the place name:

    China—Politics and government—1912-1928