Subject Access to Digital Collections in Scholars' Bank:
Background and Rationale for UO Decisions
12/30/05cgh

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. Without the supporting system functionality, 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 and its utility is questionable.

DSpace software, used to build Scholars' Bank at the UO, does not support the MARC format. Instead, fields are mapped to Dublin Core (DC). As part of the default submission template, subject information (labeled "keyword") is mapped to DC Subject, whether the field contains a personal name, corporate name, geographic name, time period, or topical term. Although it is possible to map subject terms to a particular encoding schema (such as LCSH), such mapping causes the term to be hidden from public display while still remaining searchable. Additionally, DSpace software provides no support for having a controlled list of terms for input into a subject field, nor does it provide any mechanism for cross references from variant terms. Lists of controlled terms must be consulted outside of the DSpace framework. There is also no way, without resorting to querying the underlying database, of determining terms that have been used as keywords or subjects within any DSpace collection.

It should also be noted that the DSpace software follows a decentralized submission model, whereby individual authors or designated representatives submit files and fill in the metadata (author, title, keyword, etc.) on the submission template. Because the developers of the software

expected each community of users to have its own standards for the type of metadata to be supplied they built in no support for controlled vocabularies.

General policies

There are two general policies for subjects (keyword) that are followed across collections in Scholars' Bank. The first policy is to input only one term or one subject string per input box. Two or more distinct terms should be entered in separate input boxes. The second policy is to spell words correctly.

Because of the reduced functionality of DSpace software when compared to a library catalog or even other digital content management systems such as CONTENTdm, the UO Libraries has no set input standards for subject access to materials in Scholars' Bank. Decisions are made on a collection-by-collection basis. Factors influencing the decision are:

- who is submitting files and filling in metadata for a collection (library staff versus individual authors or their designated representatives)
- whether there is any controlled list of terms that is normally used by the community
- whether full-text indexing will make up for the lack of content analysis as represented by the terms supplied in the keyword (subject) field
- ease of input and time needed to research and locate keywords

Although Scholars' Bank is registered with several open access registries and the metadata from it is harvested or indexed for use outside of the local repository, there are no shared input standards for open access repositories. This is very different from our library catalog records which are contributed to the OCLC database for worldwide use where we are expected to follow strict input standards. Because Scholars' Bank records are used very differently from online catalog records, we are comfortable following less rigid input standards for keyword (subject) fields in Scholars' Bank. This approach will be reviewed as system functionality improves or as user expectations demand it.

Topical terms

LCSH

For collections where MDLS or other library staff are submitting files and entering metadata for those files, the decision may sometimes be made to use Library of Congress forms of subject headings. In those cases, the LCSH subject string will be entered as it would appear in the public catalog but without any MARC coding. For instance, a heading appears in the public display of a catalog record with two dashes separating the distinct subfields of a subject heading. In Scholars' Bank, the term would be entered in the keyword (subject) field, complete with dashes. Subfields $x, $v, $y, and $z will be retained but will be translated into two dashes. This can be easily accomplished by copying and pasting from a catalog record. For example:

United States - - Foreign relations - - China.
American poetry - - 20th century.

Because DSpace does not provide any mechanism for reviewing terms or for properly coding and displaying terms from different schema, the use of LCSH subject strings can be based on searches of the public catalog with terms being copied and pasted in. In the current environment, no attempt will be made to review or clean up these terms.

TGM, AAT, etc.

Sometimes other source vocabularies are used for supplying keywords for particular Scholars' Bank collections. There is currently no way to map these to the encoding schema and have the terms display properly. In the current environment, no attempt will be made to review or clean up these terms.

Local vocabularies or free text

Individual authors or designated collection representatives who submit files to Scholars' Bank may supply any keywords that they consider appropriate. In the current environment, no attempt will be made to review or clean up these terms.

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 and entered in the keyword field in the submission template. MDLS staff entering names as subjects in Scholars' Bank collections 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 the keyword field.

Place names

Forms of place names may be researched in LCNAF, GNIS, or the Columbia Gazetteer, if desired. When input by MDLS staff, forms of place names most often follow AACR2 guidelines. However, no consistent input standards are applied for the forms of place names across collections in Scholars' Bank. In the current environment, no attempt will be made to review or clean up these terms.

Clean up and review

The only clean-up and review of metadata supplied in the keyword field is done by the metadata reviewer who has been authorized for each collection. At this writing (Dec. 2005) all metadata reviewers are library faculty – either in MDLS or the Document Center.