A Faceted Approach to Subject Data in the Dublin Core Metadata Record

Lois Mai Chan
Eric Childress
Rebecca Dean
Edward T. O'Neill
Diane Vizine-Goetz

SUMMARY. The enormous volume and rapid growth of resources available on the World Wide Web and the emergence of numerous metadata schemes have spurred a re-examination of the way subject data is to be provided for Web resources efficiently and effectively. For the Dublin Core metadata record, a new approach to subject vocabulary is being investigated. This new method, called FAST (Faceted Application of Subject Terminology), is based on the existing vocabulary in *Library of Congress Subject Headings* (LC), but applied with a simpler syntax than that currently used by libraries according to Library of Congress application policies. In the FAST system, non-topical (i.e., geographic, chronological, and form) data are separated from topical data and placed in different elements provided in the Dublin Core metadata record. *[Article copies*

Lois Mai Chan is Professor, School of Library and Information Science, University of Kentucky (e-mail: loischan@pop.uky.edu).

Eric Childress is Senior Product Support Specialist, Library Resources Division, OCLC Online Computer Library Center, Inc. (e-mail: eric_childress@oclc.org).

Rebecca Dean is Manager, Authority Control Section, OCLC Online Computer Library Center, Inc. (e-mail: rebecca_dean@oclc.org).

Edward T. O'Neill is Research Scientist, Office of Research, OCLC Online Computer Library Center, Inc. (e-mail: oneill@oclc.org).

Diane Vizine-Goetz is Research Scientist, Office of Research, OCLC Online Computer Library Center, Inc. (e-mail: vizine@oclc.org).

[Haworth co-indexing entry note]: "A Faceted Approach to Subject Data in the Dublin Core Metadata Record." Chan et al. Co-published simultaneously in *Journal of Internet Cataloging* (The Haworth Information Press, an imprint of The Haworth Press, Inc.) Vol. 4, No. 1/2, 2001, pp. 35-47; and: *CORC: New Tools and Possibilities for Cooperative Electronic Resource Description* (ed: Karen Calhoun, and John J. Riemer) The Haworth Information Press, an imprint of The Haworth Press, Inc., 2001, pp. 35-47. Single or multiple copies of this article are available for a fee from The Haworth Document Delivery Service [1-800-342-9678, 9:00 a.m. - 5:00 p.m. (EST). E-mail address: getinfo@haworthpressinc.com].

available for a fee from The Haworth Document Delivery Service: 1-800-342-9678. E-mail address: <getinfo@haworthpressinc.com> Website: <http:// www.HaworthPress.com>]

KEYWORDS. Dublin Core, FAST (Faceted Application of Subject Terminology), LCSH (Library of Congress Subject Headings), metadata, subject analysis

INTRODUCTION

The rapid growth of the Internet and the number of electronic resources now available on the Web have necessitated a re-thinking and re-assessment of the way in which information resources are described and made accessible. The proliferation of metadata schemas and the eagerness with which they are embraced, even before they are fully developed and refined, by various communities are a reflection of the urgent need for new approaches to knowledge organization and maintenance.

Many of the metadata schemas have been developed with specific communities or specific types of resources in mind. Among these schemas, The Dublin Core Initiative is the broadest in scope, attempting to meet the needs of a wide variety of user communities and covering resources in all subject areas and all types of electronic resources.

PREMISE OF DUBLIN CORE

Since its inception, Dublin Core has attracted the attention of various user communities, including libraries, archives, museums, and other information providers. In order to accommodate a wide range of users, the premise of the Dublin Core has been stated in the following terms:

> The Dublin Core is intended to be usable by non-catalogers as well as resource description specialists. (The Dublin Core: A Simple Content Description 1998)

In other words, the intention of Dublin Core is to have a schema that mediates between the unstructured approach to Web resources and the highly sophisticated resource description schemas such as AACR2R/MARC.

The following characteristics distinguish the Dublin Core as a prominent candidate for description of electronic resource: simplicity, semantic inter-

operability, international consensus, and flexibility (The Dublin Core: A Simple Content Description 1998). To achieve these goals, fifteen elements considered to be essential for the identification and description of Web resources have been defined for the Dublin Core.

One of the fifteen elements in the Dublin Core is designated as SUBJECT (Dublin Core Metadata Element Set: Reference Description 1999):

SUBJECT

> ***The topic of the content of the resource.*** *Typically, a subject will be expressed as keywords, key phrases or classification codes that describe a topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme.*

In addition to SUBJECT, a number of other elements in the Dublin Core also contain subject-related or content-related data:

TITLE

DESCRIPTION

TYPE

LANGUAGE

COVERAGE

The definitions of these subject-related elements are set forth in the element-set (Dublin Core Metadata Element Set: Reference Description 1999):

DESCRIPTION

> ***An account of the content of the resource.*** *Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.*

TYPE

> ***The nature or genre of the content of the resource.*** *Type includes terms describing general categories, functions, genres, or aggregation levels for content. Recommended best practice is to select a value from a controlled vocabulary.*

LANGUAGE

> ***A language of the intellectual content of the resource.*** *Recommended best practice for the values of the Language element is defined by RFC 1766, which includes a two-letter Language Code (taken from the ISO 639 standard) followed, optionally, by a two-letter Country Code (taken from the ISO 3166 standard). For example, 'en' for English, 'fr' for French, or 'en-uk' for English used in the United Kingdom.*

COVERAGE

> ***The extent or scope of the content of the resource.*** *Coverage will typically include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity). Recommended best practice is to select a value from a controlled vocabulary (for example, the Thesaurus of Geographic Names) and that, where appropriate, named places or time periods be used in preference to numeric identifiers such as sets of coordinates or date ranges.*

The element *TYPE* constitutes more or less what has been considered traditionally as "form data." Taken together, these content-related elements imply a post-coordinate, faceted approach to content representation. In other words, form, language, place, and time are separate from topical representation in the SUBJECT element. Regrouped, these elements fall into the following categories:

A. Topical description–

SUBJECT

TITLE

DESCRIPTION

B. Form data–

TYPE

C. Language data–

LANGUAGE

D. Spatial or temporal data–

COVERAGE

The Dublin Core has been designed to be extremely flexible; none of the elements is mandatory, and every element is repeatable. In other words, even though there are many elements relating to subject matter, a particular implementation may choose to use all of them or only the *SUBJECT* element for representing content-related data. Each implementation project or agency must define the scope of each element and determine its own policy with regard to the use of these elements.

With regard to the development or adoption of an implementation policy, it is appropriate to consider first the nature of the Web environment for which the Dublin Core has been designed and the functional requirements of a subject access schema for Web resources.

Two of the primary characteristics of Dublin Core Metadata (The Dublin Core: A Simple Content Description 1998) mentioned earlier are particularly pertinent with regard to subject data: simplicity and semantic interoperability. Simplicity refers to the usability by non-catalogers, specifically, to allow the creation of metadata records by persons not necessarily trained in sophisticated methods of bibliographic control. The reason for semantic interoperability is to enable users to search across discipline boundaries and, desirably, also across information retrieval and storage systems.

In keeping with the premises of the Dublin Core, a subject vocabulary suitable for the Web environment has the following functional requirements:

- It should be simple in structure (i.e., easy to assign and use) and easy to maintain;
- It should provide optimal access points;
- It should be flexible and interoperable across disciplines and in various knowledge discovery and access environments, not the least among which is the OPAC.

PREVIOUS EFFORTS TO REVISE LIBRARY OF CONGRESS SUBJECT HEADINGS (LCSH)

The complexity of LCSH has prompted several simplification attempts. Among these, the Subject Subdivisions Conference, also known as the Airlie Conference (The Future of Subdivisions 1992), attempted to simplify the application of LCSH subdivisions. At that conference, many of the problems associated with LCSH subdivision practice were identified and a series of recommendations were made, including the following:

- A standard order of subdivision (topical, geographic, chronological, and form) should be used for topical headings;
- The use of free-floating subdivisions should be expanded;

- Chronological subdivisions should reflect the actual time period covered in the work.

Since the Airlie Conference, the Library of Congress has embarked on a series of efforts to simplify subdivision practice. Nonetheless, the pre-coordinated string remains.

In 1997, an ALCTS/SAC/Subcommittee on Metadata and Subject Analysis was established with the following charge: *Identify and study the major issues surrounding the use of metadata in the subject analysis and classification of digital resources. Provide discussion forums and programs relevant to these issues.* The Subcommittee did an excellent job of analyzing the needs for the subject analysis of digital resources. It started with the assumption that the schema must be simple, easy to apply, intuitive, scalable, logical, and appropriate to the specific discipline of implementation. Unlike the Airlie Conference, the Subcommittee focused on identifying different approaches to subject data rather than on enhancing LCSH.

The Subcommittee concluded that using a mixture of keywords and controlled vocabulary is the most viable approach for digital resources. The potential sources of controlled vocabulary include:

- Use existing schema(s);
- Adapt or modify existing schema(s);
- Develop new schema(s).

For a general vocabulary covering all subjects, members of the Subcommittee considered two options: *(1) Using LCSH subject strings, if possible (i.e., if time and trained personnel are available), particularly in the OPAC environment or (2) Breaking up LCSH strings to topic, place, period, language, etc., and using other Dublin Core elements (type, coverage) in addition to the SUBJECT element.* They recommended the second option, the use of *separate elements for form, chronology, type, time, and space, particularly in situations where non-catalogers are involved in the creation of metadata records.*

IMPLEMENTING SUBJECT ELEMENT IN THE CORC PROJECT

The CORC (Cooperative Online Resource Catalog) Project, initiated by OCLC in 1998, seeks to test an experimental model for the implementation of the Dublin Core, which may eventually develop into a working tool. The subject analysis requirements for CORC are very similar to those identified by the ALCTS/SAC/Subcommittee but with additional emphasis on: (1) compatibility between any new schema and LCSH, and (2) amenability to automated authority control.

Two key decisions are required to create a new subject schema: defining the semantics (the choice of vocabulary) and the syntax (pre-coordination vs. post-coordination). Regarding the semantics, OCLC decided to retain LCSH vocabulary. CORC participants are expected to continue to apply LCSH, even if a new schema becomes available and widely accepted. By adapting the LCSH vocabulary, the compatibility with LCSH is retained. As a subject vocabulary, LCSH offers several advantages:

- It is a rich vocabulary covering all subject areas, easily the largest general indexing vocabulary in the English language;
- There is synonym and homograph control;
- It contains rich links (cross references) among terms;
- LCSH is a de facto universal controlled vocabulary and has been translated or adapted as a model for developing subject heading systems by many countries around the world;
- It is compatible with subject data in MARC records;
- With a common vocabulary, automated conversion of LCSH to the new schema is possible;
- The cost of maintaining the new schema is minimized since many of the changes to LCSH can be incorporated into the new schema.

While LCSH has served users of OPACs long and well, its complex pre-coordinate syntax poses significant disadvantages:

- Because of its complex syntax and rules, the application of LCSH requires highly trained personnel;
- Subject heading strings in bibliographic or metadata records are costly to maintain;
- The syntax of LCSH is not compatible with most other controlled vocabularies;
- It is not amenable to search engines outside of the OPAC environment, particularly current Web search engines; and,
- Due to the complex rules for constructing headings, authority control is of limited effectiveness.

While the rich vocabulary and semantic relationships in LCSH provide subject access beyond the capabilities of keywords, its complex syntax presents a stumbling block and runs counter to the basic premises of simplicity and semantic operability of the Dublin Core. The preferred solution is to devise a simplified syntax using the LCSH vocabulary. The resulting schema would have a controlled vocabulary based on the terminology and relationships already established in LCSH but structured with a different syntax and applied with different policies and procedures that are more inclined towards

post-coordination. One of the advantages of such an approach is that, by separating syntax from semantics, the syntax can be simplified while the richness of vocabulary in LCSH is retained, making the schema easier to use and maintain.

The central issue involving the syntax of a controlled vocabulary is pre-coordination vs. post-coordination. Both have precedence in cataloging and indexing practice. Subject vocabularies used in MARC records are typically pre-coordinated subject heading strings, while controlled vocabularies used in online databases are mostly single-concept descriptors, relying on post-coordination for complex subjects. The structure of the elements in the Dublin Core, as discussed earlier, implies or at least allows a faceted, post-coordinate approach. For the sake of simplicity and semantic interoperability, the post-coordinate approach is more in line with the basic premises and characteristics of the Dublin Core. It is in keeping with the primary intent of the Dublin Core to make it "usable by non-catalogers as well as by those with experience with formal resource description models."

Although there is an LCSH authority file that contains established headings, its primary coverage is for the roots of constructed subject headings. Only about three percent of all of the topical headings in WorldCat match authority records–the remaining headings were formed by subdividing the established forms. The rules for constructing a heading are complex–the *Subject Cataloging Manual: Subject Headings* (*Subject Cataloging Manual* 1996) contains four volumes of complex guidelines. The correct construction of subject headings requires extensive education–typically at least two graduate level courses–and years of experience. Since many CORC users are not expected to be skilled subject catalogers, automated authority control must play a significant role to ensure the quality of the subjects assigned. The application rules could be greatly simplified by fully establishing all headings in the new schema, thus eliminating most of the rules for heading construction and greatly simplifying authority control.

THE FAST SUBJECT SCHEMA

After analyzing the requirements of the Dublin Core and reviewing the previous attempts to improve LCSH or to provide other schema for metadata, OCLC is developing a Faceted Application of Subject Terminology (FAST) schema derived from the Library of Congress Subject Headings. The FAST schema is:

- Based on the LCSH vocabulary;
- Designed for an online environment;
- A post-coordinated faceted vocabulary;

- Usable by people with minimal training and experience, and
- Compatible with authority control.

Facets

FAST is being developed in two phases. The first phase includes the development of facets based on the vocabulary found in LCSH topical and geographic headings. The sources of the four distinct types of subfields (facets) that are used in topical and geographic headings are shown in Table 1. Initially, FAST will be limited to these four facets. In a later phase of development, it is anticipated that additional facets will be added for personal names, corporate names, conference/meetings, uniform titles and name-title entries. With the exception of the period facet, all FAST headings will be established in an authority file.

Topical headings will consist of topical main headings and their corresponding general subdivisions and general subdivisions from geographic headings. Period subdivisions with topical aspects will be considered to be both general and period subdivisions for the purpose of FAST. The FAST topical headings look very similar to the established form of LCSH topical headings with the exception that all free-floating topical subdivisions will be part of the established form of the heading and all multiple subdivisions will be expanded. However, rather than establishing all possible combinations, only those that have actually been used will be established. For example, headings based on the multiple

Abortion–Religious aspects–Buddhism, [Christianity, etc.]

will be established only with the religions used–not with all known religions.

TABLE 1. Facets

Facet	Source	MARC codes
Topical	Topical term and general subdivisions from topical headings, the general subdivisions from geographic headings, and period subdivisions which have topical aspects.	650 a & x; 651 x; selected 650 & 651 y
Geographic	Name from geographic headings and the geographic subdivisions from topical headings	650 z; 651 a & z
Period	Chronological subdivisions	650 y; 651 y
Form	Form subdivisions	*650 v; 651 v*

All geographic names will be established and used in indirect order, **Ohio–Columbus** rather than in direct order, **Columbus (Ohio)**. In LCSH, names used as main headings are entered in direct order but when used as subdivisions they appear in indirect form. As in LCSH, a hierarchical name structure will be used. However, unlike LCSH, the first level geographic names will be limited to names from the *USMARC Code List for Geographic Areas* table (USMARC Code List 1998). The number of names that can be used as first level entries in LCSH is far greater than the number of names in the *Geographic Area Codes* table. For example, LC has established some relatively obscure names like **Morgan line** (sh 85-87259) as first level entries. Since the Morgan line–the boundary line drawn from Trieste to the Austrian frontier dividing Italy from Yugoslavia during the post-War occupation–spans national boundaries, it does not require qualification in LCSH. However, in FAST, it would be entered indirectly under Europe, the smallest first level area in which it is fully contained. The resulting FAST entry would be **Europe–Morgan line**. Associating each first level entry with a Geographic Area Code will facilitate the development of search features that use the hierarchical structure of geographic names in FAST. A typical portion of the geographic area code table is shown in Table 2.

The difficulty of working with geographic names can be illustrated with an example: a search of WorldCat for the name Charlevoix produces results that include the use of the name as a city, a lake, a county in Michigan, and two counties in Québec. A WorldCat search identified the 47 different geographic entries shown in alphabetical order in the left column in Figure 1. However, there are actually only 19 distinct geographic entities as shown in the right column. Many of the entities, such as **Charlevoix (Mich.)** and **Michigan–Charlevoix**, result from using different forms of the same name to generate direct and indirect entries. Others, such as **Québec (Province)–Charlevoix** and **Québec–Charlevoix**, result from inconsistent qualifi-

TABLE 2. Geographic Areas Codes

Norway [e-no]	**Oceania [po]**
Norwegian Sea	UF
Assigned code:	Oceanica
[In]	Oceania, French
North Atlantic Ocean	USE
Nova Scotia [n-cn-ns]	French Polynesia
Nyasaland	Oceanica
USE	USE
Malawi	Oceania
Ocean Island (Kiribati)	**Ohio [n-us-oh]**
USE	**Ohio River [n-uso]**
Banaba	

FIGURE 1. LCSH vs. FAST Geographic Names

Geographic Names from LCSH	FAST Geographic Names
Charlevoix (Mich.)	Michigan–Charlevoix
Charlevoix (Quebec)	Michigan–Charlevoix County
Charlevoix County (Mich.)	Michigan–Charlevoix County–Beaver Island
Charlevoix County (Quebec)	Michigan–Charlevoix County–Deer Creek Watershed
Charlevoix Harbor (Mich.)	Michigan–Charlevoix County–Holy Island
Charlevoix Region (Mich.)	Michigan–Charlevoix County–Horton Creek
Charlevoix Region (Quebec)	Michigan–Charlevoix County–Horton Creek Marsh
Charlevoix Site (Mich.)	Michigan–Charlevoix County–Marion
Charlevoix, Lake (Mich.)	Michigan–Charlevoix County–O'Neill Site
Charlevoix-Est (Quebec : Regional County Municipality)	Michigan–Charlevoix County–Peaine Township
	Michigan–Charlevoix County–St. James Township
Charlevoix-Est (Quebec)	Michigan–Charlevoix Harbor
Charlevoix-Est County (Que.)	Michigan–Charlevoix Region
Charlevoix-Est County (Quebec)	Michigan–Charlevoix Site
Charlevoix-Ouest (Quebec)	Michigan–Lake Charlevoix
Charlevoix-Ouest County (Que.)	Quebec–Charlevoix Region
Charlevoix-Ouest County (Quebec)	Quebec–Charlevoix-Est
Clermont (Charlevoix-Est, Quebec)	Quebec–Charlevoix-Est–Clermont
Deer Creek Watershed (Charlevoix County, Mich.)	Quebec–Charlevoix-Ouest
Holy Island (Charlevoix County, Mich.)	
Horton Creek (Charlevoix County, Mich.)	
Horton Creek Marsh (Charlevoix County, Mich.)	
Lake Charlevoix (Mich.)	
Lake Charlevoix (Michigan)	
Marion (Charlevoix County, Mich.)	
Michigan–Beaver Island (Charlevoix County)	
Michigan–Charlevoix	
Michigan–Charlevoix County	
Michigan–Charlevoix Region	
Michigan–Charlevoix, Lake	
Michigan–Deer Creek Watershed (Charlevoix County)	
Michigan–Horton Creek (Charlevoix County)	
Michigan–Lake Charlevoix	
Michigan–Marion (Charlevoix County)	
Michigan–Peaine Township (Charlevoix County)	
Michigan–St. James Township (Charlevoix County)	
O'Neill Site, Charlevoix County (Mich.)	
Quebec (Province)–Charlevoix	
Quebec (Province)–Charlevoix Co.	
Quebec (Province)–Charlevoix East	
Quebec (Province)–Charlevoix Region	
Quebec (Province)–Charlevoix West	
Quebec (Province)–Charlevoix-Est	
Quebec (Province)–Charlevoix-Est (Regional County Municipality)	
Quebec (Province)–Charlevoix-Ouest	
Quebec–Charlevoix Region	
Quebec–Charlevoix-Est	
Quebec–Charlevoix-Ouest	

cation. Abbreviations and inverted forms are other common sources of duplicate names for the same entity. Furthermore, a bigger problem is the difficulty in displaying the direct and indirect forms of the names together. Without reformatting the headings, logical clustering is impossible.

In FAST, all geographic names are represented in indirect order. There is no limit on the number of levels, although the need for more than three levels appears rare. Qualifiers will only be used to identify the type of geographic name (Kingdom, Satellite, Duchy, Princely State, etc.). It is expected that this simplified notation will lead to clearer displays as seen in the right column of Figure 1.

Form subdivisions will be treated as another distinct facet. The initial set of form headings will be identified by extracting form subdivisions from LCSH topical and geographic headings. Form subdivisions are currently coded both *x* and *v* in LCSH headings in WorldCat. Those coded *x* will be algorithmically identified and re-coded as *v*. The algorithm, developed by OCLC, for identifying form headings will be described in detail in a forthcoming paper.

Period headings for FAST will follow the practice recommended at the Airlie Conference: Chronological headings will reflect the actual time period of coverage for the resource. All period headings will be expressed as either a single numeric date or as a date range. For example, the default time period associated with the period, **Wars of the Huguenots, 1562-1598**, would be 1562 to 1598. However, in reducing this subdivision to simply a date range, the name of the war is lost. To prevent this loss of information, period subdivisions with topical aspects will also be treated as general subdivisions under the main heading. For example, the subdivision **Wars of the Huguenots, 1562-1598** would be treated as if it had been entered as the general subdivision **Wars of the Huguenots, 1562-1598** and the period subdivision **1562-1598**, ensuring that both the chronological and topical aspects are retained in the appropriate facet. Further, since a period heading should reflect the actual time period covered, for a work covering only a single battle, e.g., one that occurred in 1565, the period heading would be limited to that single year.

Creation of FAST Authority Files

The initial FAST authority files will be built by faceting LCSH headings from WorldCat. For example the following LCSH heading,

> ***France $x History $y Wars of the Huguenots, 1562-1598 $v Juvenile literature***

would result in the following FAST headings:

Topical: **History–Wars of the Huguenots, 1562-1598**

Geographic: **France**

Period: **1562-1598**

Form: **Juvenile literature.**

A file containing all unique Library of Congress topical and geographic subject headings in WorldCat has been created. This file contains 6,912,980 unique topical and 1,471,023 geographic headings, representing over 50 million individual subject heading assignments. These headings will be faceted to create the initial versions of the FAST topical, geographic, and form authority files.

These initial versions of the FAST authorities will undergo extensive validation to minimize the number of erroneous entries. The entries remaining after this validation step will be established as FAST subject headings. The final step in creating the FAST authority files will be the addition of cross-references, notes, and other similar information necessary to convert established heading to an authority record.

CONCLUSION

Providing subject data in the CORC metadata record presents both a challenge and an opportunity to explore new approaches to subject analysis and access to electronic resources and to test the viability of an existing subject vocabulary (LCSH in this case) in the Web environment. The purpose of adapting the LCSH with a simplified syntax in CORC is to retain the very rich vocabulary of LCSH while making the schema easier to maintain, apply, and use.

REFERENCES

The Dublin Core: A Simple Content Description Model for Electronic Resources: Metadata for Electronic Resources. (1998). (http://purl.org/DC/index.htm)

Dublin Core Metadata Element Set: Reference Description. (1999). (http://purl.oclc.org/dc/documents/rec-dces-19990702.htm)

The Future of Subdivisions in the Library of Congress Subject Headings System: Report from the Subject Subdivisions Conference Sponsored by the Library of Congress, May 9-12, 1991, edited by Martha O'Hara Conway. Washington, DC: Cataloging Distribution Service, Library of Congress, 1992.

Subject Cataloging Manual: Subject Headings. 5th ed. Prepared by the Cataloging Policy and Support Office, Library of Congress. Washington, D.C.: Cataloging Distribution Service, Library of Congress, 1996-.

USMARC Code List for Geographic Areas. Prepared by the Library of Congress Network Development and MARC Standards Office. Web Version of 1998 Edition. (http://lcweb.loc.gov/marc/geoareas/)