Michael Buckland
Ryan Shaw

4W Vocabulary Mapping Across Diverse Reference Genres

Abstract

This paper examines three themes in the design of search support services: linking different genres of reference resources (e.g. bibliographies, biographical dictionaries, catalogs, encyclopedias, place name gazetteers); the division of vocabularies by facet (e.g. What, Where, When, and Who); and mapping between both similar and dissimilar vocabularies. Different vocabularies within a facet can be used in conjunction, e.g. a place name combined with spatial coordinates for Where. In practice, vocabularies of different facets are used in combination in the representation or description of complex topics. Rich opportunities arise from mapping across vocabularies of dissimilar reference genres to recreate the amenities of a reference library. In a network environment, in which vocabulary control cannot be imposed, semantic correspondence across diverse vocabularies is a challenge and an opportunity.

Introduction

The move away from cards and print to an online graphical user interface has important consequences for search support. This paper draws on studies at Berkeley to examine three different aspects of search support: overcoming the traditional separation between different genres of reference resource; the dividing of vocabularies by facet (e.g. What, Where, When, and Who); and mapping between different and dissimilar vocabularies. Important relationships between these three themes are noted.

Our starting point is the principle that understanding requires a knowledge of context and our approach to providing learners with contextual knowledge is to try to recreate the functionality of a traditional reference library in a digital environment.

Different Genres of Reference Resource

Several different forms of reference work have evolved for specialized purposes: Bibliographies, biographical dictionaries, chronologies, encyclopedias, library catalogs, place name gazetteers, and others. In a print environment any entry in any of these resources might cite any other, but, except for the library catalog, rarely tells one where a copy can be found. With the technology of card and paper these resources are not physically connected.

Sixty years ago, Ray Swank (1951) tried to combine the subject indexing of bibliographies with the locations listed in library catalogs, but this was not practical until the 1980s when digital technology became sufficiently available and affordable for records in published bibliographies to be linked to call numbers in local library records (Buckland 1988).

A digital, networked environment invites connectivity. Authority records on cards *mention* authoritative sources; authority records online can *connect* in real time with authoritative, explanatory resources, e.g. a place name authority record can link to a place name gazetteer. A link to a specialized reference work is likely to provide more and more up-to-date information than would be provided in an authority record. A well-formed place name gazetteer will not only couple each place name with spatial coordinates (latitude and longitude) for that place, but will also include a geographic

description code ("feature type") indicating what category of place it is (lake, city, mountain, etc.). In turn, the gazetteer could contain links to more extensive descriptions of the place in encyclopedias, pictures of it, map displays, socio-economic data series, and mentions in literary works.

We may think of representation in a knowledge organization system as different from description (as in an encyclopedia), but that distinction is not helpful here. The important point is that the historic, physical separation of the catalog from reference works need not and should not continue.

Divided Vocabularies

A general purpose knowledge organization system, such as a list of subject headings or a library classification, will include all kinds of topics: activities, animals, events, institutions, persons, places, and so on. However, it can be more effective and probably more efficient to separate out categories of headings for separate treatment. In divided library card catalogs, personal names for the subjects of biographies are ordinarily separated from other subject headings and interfiled with the records for authors to constitute a *name catalog* for persons, both as authors and as subjects.

A graphical user interface allows the display of different kinds of relationships. For example, who, now, would want an online catalog without a map display showing where each place is and its geographical relationship to other places? This is feasible if the name of the place can be combined with its spatial coordinates of latitude and longitude. (Buckland, Chen, Gey, Larson, Mostern and Petras 2007).

We have been working with the four facets of WHAT (topic), WHERE (place), WHEN (time), and WHO (person) because these facets are different in kind and each has distinctive characteristics and display requirements: Semantic syndetic structure for topics; map displays and complex spatial relationships for places (e.g. Larson & Frontiera 2004); and time-lines and chronologies for time periods. Displays of family trees and other interpersonal relationships are desirable in biographies. Geotemporal arrangement of biographies into "life paths" or collective analysis as prosopography may be useful. There are other possibilities: We are currently considering the separation of named events from other topics (Lancaster, Buckland and Shaw 2007).

Divided Vocabularies Combined in Representations

Facets are, in principle, different in their nature, but topics often cannot be expressed in terms of a single facet. For example, 005.912=112.2(075)(410) representing "German language primers on office management in the U.K." in the Universal Decimal Classification, is composed of a term from each of four facet vocabularies. Library of Congress Subject Headings (LCSH) routinely occur as compound representations in which the main heading is qualified by chronological and/or geographical subdivisions (e.g. Architecture – Japan – Meiji period, 1868-1912). (For discussion of LCSH and faceted interfaces see McGrath (2007)). Of course, textual descriptions in encyclopedia articles or elsewhere will use all kinds of vocabulary.

As already noted a well-formed place name gazetteer will not only couple each place name with the latitude and longitude for that place, but will also included a geographic description code (commonly called "feature type") indicating what category of place it

is (lake, city, mountain, etc.). Thus each WHERE is also categorized as a WHAT and, since names and boundaries are often unstable, should also be encoded for WHEN.

The treatment of time is less well-developed than place. In ordinary discourse and in metadata it is common to express time in geographical and cultural terms by using named events adjectivally to denote period, e.g. *Victorian* literature, a *Civil War* weapon, *Louis Quinze* furniture, and so on. LCSH chronological subdivisions ordinarily define time using periods named in terms of the political history of the topic (e.g. « Meiji period ») with calendar dates added for clarification. The use of named events to denote spans of time has more evocative connotations than mere calendar time because a cultural context is identified that might otherwise not have been recognized when relying on dates alone. Calendar time is needed, however, to map temporally between the periods of different cultures. (Petras, Larson and Buckland 2006).

Biographical texts are rich in actions, locations, dates, and other persons, but mark up standards and best practices are still generally unsatisfactory (Text Encoding Initiative 2006). We are examining the feasibility of decomposing lives into a series of events and using 4-tuple of What, Where, When, and Who to characterize each life event: an activity or event (birth, education, employment, etc.) in a place at some time, sometimes with other people (Electronic Cultural Atlas Initiative 2006).

The use of terms from different facets in compound representations and textual descriptions provides an opportunity to link any identifiable What, Where, When, or Who to an explanatory description in some other reference work. The examples given above can be briefly summarized:

Table 1. Components of examples different reference genres

WHAT (LCSH) : Topic – Geographical subdivision – Chronological subdivision
WHERE (Gazetteer) : Place name – Feature type – Latitude and longitude – When
WHEN (Chronology) : Period name – Period type – Dates – Where
WHO (Biogr. Dict.) : Personal name – Actions – Places – Dates – Other persons.

These examples provide the basis for a two-dimensional 4W array: The WHAT records have WHERE and WHEN subdivisons; the WHERE records (place name gazetteer entries) should have WHAT (feature type) and WHEN elements; WHEN records in our design for a time period directory include elements for WHAT (period type) and WHERE; and a WHO record in a biographical directory will contain multiple WHAT (action, status), WHERE, WHEN, and WHO elements. We can note a WHAT (Topic, Type, Type, Action) element in each row and also WHERE and WHEN in each. Rearranging the components in each row yields a new table in which the facets are aligned vertically.

Table 2. Two dimensional array of components of reference genres

WHAT (LCSH):	What	Where	When	
WHERE (Gazetteer):	What	Where	When	
WHEN (Chronology):	What	Where	When	
WHO (Biogr. Dict.):	What	Where	When	Who

Displayed this way, the opportunities for mapping vertically between vocabularies in reference genres become apparent. Semantic associations can be established by vertical vocabulary mapping and contextual associations can be established through horizontal associations.

Mapping Between Similar Vocabularies

Ordinarily vocabulary control is within a single vocabulary and interoperability through vocabulary mapping is between resources of a similar type, meaning at the same horizontal level in Table 1. Similar resources commonly use different knowledge organization vocabularies that are the same (or overlap) in scope but differ in form. The *Library of Congress Subject Headings* (LCSH), the Dewey Decimal Classification and the Library of Congress Classification are different and look different (e.g. for Economics: “Economics,” “330,” and “HB1”), yet they are similar in that they are general purpose systems for providing topical access. Vocabulary interoperability is ordinarily based on cross-walks defining comparable fields in two or more similar resources and mapping between terms in those fields.

Mapping Between Dissimilar Vocabularies

The attraction of linking between *different genres* of reference resource is the wider range of descriptions and so the richer context. The vertical columns Table 2 reveal a greatly expanded scope for mapping between dissimilar vocabularies. For example, the feature type codes of a place name gazetteer can be mapped to the topics in a subject heading list. We examined the 600+ Geographic Description Codes (GDC) of the US federal National Geo-Intelligence Agency in relation to the far larger *Library of Congress Subject Headings* and found that, despite stylistic differences, there was, in most cases, a identifiable equivalent. This mapping allows a connection between a category of geographical feature (e.g. a lighthouse) and literature about lighthouses. In the other direction one could go from literature to physical examples, their locations, and a map display showing where they are located. One could move, for example, from the Geographic Description Code “school” to the corresponding library catalog subject heading and on to related literature. Since a gazetteer is concerned with physical, geographical features the semantic equivalent of the geographical feature « School » in LCSH is “School buildings,” but, situationally, “Schools” (denoting schools as institutions) may be preferred. In both GDC and LCSH the option of displaying related terms would be highly desirable (Buckland, Chen, Gey, Larson, Mostern and Petras 2007). For examples of literature being related to maps see Moretti (1988).

Such mapping may seem inappropriate because subject headings are concerned with topics and feature types with physical objects and these are different in kind and, therefore, not comparable. This criticism is correct in principle, but misguided. It is precisely this difference that enables one to move from literature to physical objects and vice versa, a genuinely multimedia approach.

Why Vocabulary Mapping Is Hard

There is a widely-held assumption that search support in the Web should be provided by ontologies and that these ontologies can be made interoperable. (DeRidder 2007 provides an overview). This seems reasonable until one tries to do it and/or one

reflects on the nature of language and/or is concerned with representations of knowledge rather than of widgets and/or contemplates the size of the task. Mapping between knowledge organization vocabularies is hard for several reasons:

1. Vocabularies are descriptive. Assigning descriptive metadata (descriptors, index terms, classification codes, category codes) is a language activity even through documentary languages are more or less artificial. They are incorrigibly context-specific and inherently obsolescent (Buckland 2007) .

2. Vocabularies are cultural. The classic definition by E. B. Tylor (1871, v.1, p.1) states "Culture or Civilization, taken in its wide ethnographic sense, is that complex whole which includes knowledge, belief, art, morals, law, custom, and any other capabilities and habits acquired by man as a member of society." More recent definitions refer to integrated patterns of human knowledge, belief, and behavior, agree that culture is not genetically transmitted but learned, and position knowledge as a part of culture. It follows that knowledge organization is concerned with learned, cultural entities and it is a characteristic of what is cultural (and of what is learned) that it is subjective, more or less imprecise, and unstable. Discourse about culture tends to use archetypes for lack of clear definitions and instances often form a smooth continuum from one type to another. The implications for knowledge organization are significant. In general terms, it must mean that precise, logical representations are unavoidably Procrustean distortions. Vocabularies such as thesauri and classification schemes are valuable and necessary. Nevertheless vocabularies become increasingly problematic as their use is extended beyond a single, local application and/or are continued over time.

3. Language, being cultural, evolves within domains of discourse and it is not to be expected that any given vocabulary will correspond precisely to one homogenous community. Typically a knowledge organization vocabulary is a linguistic compromise. Ideally, there would multiple indexes to any given resource, one for each community of users, a challenge addressed by Petras (2006).

4. There is a lot mapping to be done! The whole logic of an internet environment is to greatly increase access to resources with more-or-less unfamiliar descriptive metadata, greatly increasing the need and scope for mapping between vocabularies. Mapping is best done by well-qualified experts, but it is very labor-intensive. Fortunately, statistical association techniques combined with natural language processing can provide rapid, inexpensive search term recommender services where a training set can be found (Buckland and others 1999; Gey, M. Buckland, A. Chen and R. Larson. 2001).

Conclusions

Understanding requires a knowledge of context so it is important to enable learners to use diverse reference genres to find contextual information. In a traditional, paper-based reference library the various resources are not physically linked but cite each other and are physically collocated so that it is humanly possible to move from an entry in one to an entry in another. That functionality has yet to be reconstructed in a digital

library environment (Buckland 2008), but the adoption of online graphical user interfaces allows many important developments. Dividing vocabularies by facet simplifies the task of assigning descriptive metadata, facilitates specialized displays, and constitutes a form of infrastructure (Buckland 2006). Combining divided vocabularies is necessary for representation and descriptions. Mapping across both similar and dissimilar vocabularies is a difficult but necessary part of a broadly based learning environment.

Acknowledgments. This paper draws on the work of several other people, especially Fredric Gey, Ray R. Larson and Vivien Petras. We are grateful for the support of the Institute of Museum and Library Services National Leadership Grant for Libraries, LG-02-02-0035-02; LG-02-04-0041-04; and LG-06-06-0037-06 and to both the National Endowment for the Humanities and the Institute of Museum and Library Services for an Advancing Knowledge grant PK-50027-07 (Electronic Cultural Atlas Initiative. 2002, 2004, 2006 and 2007).

References

Buckland, M. 1988. Bibliography, library records, and the redefinition of the library catalog. *Library Resources & Technical Services* 32: 299-311.

Buckland, M. 2006. Description and search: Metadata as infrastructure. *Brazilian Journal of Information Science* vol 0 (2006). http://www.portalppgci.marilia.unesp.br/bjis/

Buckland, M. 2007. Naming in the library: Marks, meaning and machines. In: *Nominalization, nomination and naming in texts*. C. Todenhagen & W. Thiele, eds. Tübingen, Germany: Stauffenburg Verlag, pp 249-260. http://people.ischool.berkeley.edu/~buckland/naminglib.pdf

Buckland, M. 2008. Reference service in the digital environment. *Library & Information Science Research* Forthcoming.

Buckland, M., A. Chen, H-M. Chen, Y. Kim, B. Lam, R. Larson, B. Norgard, and J. Purat. 1999. Mapping entry vocabulary to unfamiliar metadata vocabularies. *D-Lib Magazine* 5 http://www.dlib.org/dlib/january99/buckland/01buckland.html

Buckland, M., A. Chen, F. C. Gey, and R. R. Larson. 2006. Search across different media: Numeric data sets and text files. *Information Technology and Libraries* 25: 181-189. http://www.lita.org/ala/lita/litapublications/ital/252006/number4decembera/buckland.pdf

Buckland, M. A. Chen, F. C. Gey, R. R. Larson, R. Mostern & V. Petras. 2007. Geographic search: Catalogs, gazetteers, and maps. *College & Research Libraries* 68: 376-387. http://www.ala.org/ala/acrl/acrlpubs/crljournal/collegeresearch.cfm

Buckland, M., and L. Lancaster. 2004. Combining time, place, and topic: The Electronic Cultural Atlas Initiative. *D-Lib Magazine* 10 http://www.dlib.org/dlib/may04/buckland/05buckland.html

DeRidder, J. L. 2007. The immediate prospects for the application of ontologies in digital libraries. *Knowledge Organization* 34: 227-246.

Electronic Cultural Atlas Initiative. 2002. *Going places in the catalog: Improved geographic access*. http://ecai.org/imls2002

Electronic Cultural Atlas Initiative. 2004. *Support for the learner: What, where, when, and who*. http://ecai.org/imls2004

Electronic Cultural Atlas Initiative. 2006. *Bringing lives to light: Biography in context*. http://ecai.org/imls2006

Electronic Cultural Atlas Initiative. 2007. *Context and relationships: Ireland and Irish studies*. http://ecai.org/neh2007

Gey, F., M. Buckland, A. Chen and R. Larson. 2001. *Entry vocabulary: a technology to enhance digital search*. http://metadata.sims.berkeley.edu/papers/hlt01-final.pdf

Lancaster, L., M. Buckland, and R. Shaw. 2007. Event: Occurrence and agency in digital resources. In: *6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference. Workshop 9: Cultural Heritage on the Semantic Web, Busan, Korea. (12 Nov 2007)*. Pp. 77-82. http://www.cs.vu.nl/~laroyo/CH-SW/ISWC-wp9-proceedings.pdf

Larson, R. L. and P. Frontiera. 2004. *Spatial ranking methods for geographic information retrieval (GIR) in digital libraries*. Paper presented at the European Collaborative Digital Library Conference, 2004. http://cheshire.lib.berkeley.edu/ECDL2004_preprint.pdf

McGrath, K. 2007. Facet-based search and navigation with LCSH : Problems and opportunities. *Code4Lib Journal* 1. http://journal.code4lib.org/articles/23

Moretti, F. 1998. *Atlas of the European novel*. London : Verso.

Petras, V. 2006. *Translating dialects in search: Mapping between specialized languages of discourse and documentary languages*. Doctoral dissertation, University of California, Berkeley. http://www.sims.berkeley.edu/~vivienp/diss/vpetras-dissertation2006-official.pdf

Petras, V., R. R. Larson, and M. Buckland. 2006. Time period directories: A metadata infrastructure for placing events in temporal and geographic context. In: *Opening Information Horizons; 6th ACM/IEEE-CS Joint Conference on Digital Libraries 2006*. New York: Association for Computing Machinery, pp.151-160 http://portal.acm.org/citation.cfm?id=1141782

Swank, R. C. Subject cataloing in the subject-departmentalized library. In: *Bibliographic organization*, ed. by J. H. Shera & M. E. Egan. Chicago: University of Chicago Press, pp. 187-199. Repr. in Swank, R. C. 1981. *A unifying influence*. Metchen, NJ: Scarecrow Press, pp.177-190.

Text Encoding Initiative. 2006. *TEI: Personography Task Force.* http://www.tei-c.org/Activities/Workgroups/PERS/

Tylor, Edward Burnett. 1871. *Primitive culture*. London : Murray.