Submitted for publication. Draft of Nov 4, 2006.

Description and Search: Metadata as Infrastructure

Michael K. Buckland
School of Information,
University of California, Berkeley, CA 94720-4600

Abstract
The first and original use of metadata is for describing documents. XML, the Dublin Core and MARC library catalog records are examples. The name "metadata" (beyond or with data) and the popular definition "data about data" is based on this use. A second use of metadata is to form organizing structures by means of which documents can be arranged. These structures can be used both to search for individual documents and also to identify patterns within a population of documents. The second role of metadata involves an inversion of the relationship between document and metadata. These structures can be considered infrastructure.

The First Purpose of Metadata: Description

The term "metadata" is used to denote "data about data" and its first and original purpose is to describe documents. (Here we do not distinguish between *data* and *documents*). There are different kinds of descriptive metadata: technical (to describe format, encoding standards, etc.); administrative (to describe intellectual property rights, conditions of access, etc.); and content (subject matter, scope, authorship, etc.). These descriptions characterize and explain the data. Metadata helps one to understand what the data is and how to make use of it (Caplan 2003, Haynes 2004).

Metadata has two components: A format and a set of values. XML, the Dublin Core and MARC library catalog records are well-known formats and are associated with specific standards for specifying the kinds of descriptions that may be used with them. Description can be very useful, even if idiosyncratic terminology is used. Almost any description is better than none. However, it is always strongly recommended that descriptive metadata follow standardized forms, e.g. using a standard format and widely used terminology. The use of standardized formats for storing and displaying makes use of metadata easier. The use of standard vocabularies has the advantage of consistency and aids understanding.

All description is a language activity even if an artificial notation, such as the Dewey Decimal Classification system, is used. Description is always and necessarily culturally-based because descriptions are based on the concepts, definitions, and understandings that have developed in a community.

When you browse documents, especially digital documents, you are likely to examine descriptive metadata in order to understand what kind of document it is, what it is about, and how to use it. This process resembles the way one can look at the cover of a book to help assess the text within.

Infrastructure

"Infrastructure" is a collective term for the subordinate parts of an undertaking. The word was initially used to refer to fixed resources used for transportation and military operations and has been gradually extended to include services ancillary to, or in support of, the performance of a central task. Minimally, travel by train requires tracks, a locomotive, and wagons, but an effective and reliable railroad service depends on other auxiliary resources: systems for signalling, ticketing, and communicating between stations, fuel depots, a management hierarchy, the publication of timetables, and so on. The collective name for these auxiliary resources is "infrastructure."

However, both the term "infrastructure" and the fashionable variant "cyberinfrastructure" (National Science Foundation 2003) are somewhat problematic. Infrastructure is always some kind of structure, but which structures should be considered *infra*structure is not always clear and is situational. A modern bank depends on data processing services to provide its banking services and this computing support is considered part of the bank's infrastructure. For the computing services industry, auxiliary resources -- the infrastructure -- includes reliable banking services for handling payments. So banking services are, in turn, part of the infrastructure of the computer services industry.

Standards and protocols form an intangible form of infrastructure with very tangible consequences and, since infrastructure is considered to be the environment of support that enables and empowers, an argument can be made that the social conventions and mentalities (the structures of thought discussed in Michel Foucault's *The Order of Things* (1970)) could be considered a form of infrastructure (Day 2006).

The second use of metadata

Thinking of metadata as a means of describing individual documents reflects only one of the two roles of metadata. The second use of metadata is different: It occurs when you *start* with the metadata rather than the data, with the description rather than the document. This occurs when you search in a catalog or browse in any index.

This second use of metadata is for search and selection. In a digital environment, that is, so far, composed mainly of text resources, it is common and convenient to use textual queries and to search for the occurence of text fragments in the available documents, as in web search engines. So, a search for the topic "Mouse" is expressed as the character string "m o u s e" and every document containing that sequence of characters will be retrieved as "relevant" whether it is really about this kind of mammal, or refers to the computer device, or arises in other, figurative uses of the word. The technique of searching by text character strings works quite well but not always and not perfectly, because text resources are not entirely homogeneous. Some words have multiple meanings (polysemy, e.g. *mouse*); sometimes different words use the same character string but with other meanings (homographs, e.g. *pane* means a sheet of glass in English, but not in Portuguese); and different words may be used with the same meaning (synonyms, e.g. *cancer* and *neoplasm*).

Simple text searches break down in multilingual environments and when other kinds of resources are included, such as images, sounds, and numeric data sets. Images can be compared with each other, and sound can be compared with other sounds, but not, directly, with other media forms. One cannot use a query composed of a few pixels, or a sound, as a query in a text file.

Indexes as Structure

To establish a meaningful connection between different documents, two actions are required: First, make a connection between them, and then express the nature of the relationship between them. For example one might assign the same topic description, e.g.

-- a text is assigned a subject heading, and

-- an image is asssigned the same subject heading

The next step is to create an inverted form of this relationship so that one can go from the subject heading both to the text and also to the image. This allows a unified search of both texts and images relating to the same topic. It also allows a transverse search from a text through a subject heading to images assigned the same subject heading, or, equally, from an image to texts.:

Text ----- subject heading ------ images

In this way, two or more documents on the same topic, however diverse their format or content, can be related to each other. This process depends on having a common descriptive vocabulary to describe topics, or, at least, interoperable vocabularies.

This last manoeuver inverts the original structure. Instead of descriptions being attached to documents, the documents are attached to the descriptions. The vocabulary of the descriptions becomes central and the documents become peripheral. This inversion is clearly seen in citation indexes. When you examine books and articles, the references are peripheral, in footnotes or at the end, and are often in smaller type. But a citation index inverts that relationship. The citations themselves and the relationships between them become primary. Only when a *citation* of interest has been selected is a *document*, at the periphery, consulted.

Syndetic Structures

The relationships ("syndetic structure") between different headings in a descriptive vocabulary (preferred terms, non-preferred terms, broader, narrower, and other related terms) are well-understood in information science. But a much larger problem has been neglected. In a non-network situation you need cross-references *within* a vocabulary. The purpose of a network is to enable access to multiple different resources, and in a network environment you need cross-references *between* vocabularies. You need to know, for example, that for, say, *Automobiles* are labeled *TL 205* in the Library of Congress Classification, but as *180/280* in the US Patent Classification, as *3711* in the Standard Industrial Classification, and as *PASS MOT VEH* in the U.S. import and export statistical series.

This problem is not unknown and terminology exists: A link between different but similar indexes is called a *crosswalk*, and links between comparable headings in two or more different indexes is called a *mapping*. An outstanding example is the Unified Medical Language Sytems (UMLS) of the National Library of Medicine (2006), but, otherwise, such mappings are rarely provided and, indeed, are rather infrequently discussed. The expertly-made mappings of the UMLS system are expensive and, of course, obsolescent, but in many circumstances statistical association techniques can be used to generate useful mappings across vocabularies (Buckland, Gey & Larson 2002, Buckland, Chen, Gey & Larson, forthcoming). The general neglect of cross-vocabulary

mapping suggests that the implications of a networked environment have not been fully understood.

Personal names

Personal names are important for authorship and biographical texts. The need to differentiate between different persons with the same name and aggregating different names for the same person is well-understood in archives, libraries, museums, and elsewhere. However, the techniques for handling interpersonal relationships appears to have been rather neglected. Genealogists have experience with encoding family relationships (parent-child, spouse, etc.), but people can be related to each other in other important ways (e.g. teacher-pupil, business partner) for which techniques and terminology need further development.

Geographical Areas: Place and Space

Searching in a text environment is dominated by topical keywords or undifferentiated keywords, possibly including the names of persons, places, and institutions. However, for searching in some resources, such as socio-economic data series and photographs, it becomes important to specify geographical location reliably and exactly. “Place” is a cultural construct and this is reflected in place names, which, like topic names, are often multiple (e.g. Lisboa, Lisbon, Lisbona, Lisbonne, Lissabon), ambiguous (Galicia, Poland; Galicia, Spain), and unstable (e.g. St Petersburg became Leningrad then St Petersburg again).

Space, in contrast, is defined in physical terms of latitude and longitude, which provides descriptions that are neither ambiguous nor unstable. A large advantage of spatial coordinates is that they allow places to be shown on a map. There is, therefore, for geographical areas, a dual naming system of place and space: place names and spatial coordinates. A place name gazetteer can be considered a kind of bilingual dictionary between places and spaces. A gazetteer enables place names to be disambiguated and places to be located on a map. A well-designed gazetteer will indicate when a place name was in use, thereby supporting temporally-dynamic maps (Zerneke, Buckland & Carl 2006). (For a discussion of the use of gazetteers and map interfaces to improve searching see Buckland, Gey & Larson 2004; Buckland, Chen, Gey, Larson, Mostern & Petras, forthcoming.)

Events and Time

Events and time tend to be mutually defining. Time is calibrated by physical events and cultural epochs by cultural events. But physical events and cultural epochs are also calibrated by calendar time. In speech and in writing, we commonly mark time by reference to events, as in “after I graduated” or “before the Second World War.” This duality of events and of time resembles the duality of place and space and invites a similar approach: the use of a directory relating named events to calendar time. Associating events with dates supports the construction of time lines and chronologies in the same way that a place name gazetteer relates place names to spatial coordinates and map displays (Petras, Larson & Buckland 2006).

Biography

Although, as already noted, information scientists have effective methods for handling peoples' *names*, methods for handling the *events* in their lives are much less developed (Text Encoding Initiative Consortium 2006). One approach being investigated by the Electronic Cultural Atlas Initiative is to categorize each biographical event or life activity as a four-facet tuple of what kind of activity (topical aspect), when (temporal aspect), where (geographical aspect), and with whom (biographical aspect) (Bringing 2006). An attraction of this approach is that life events could be encoded with the terminology and methods already established, or being developed, for subject indexing, time periods, place names, and biographical dictionaries.

Infrastructural relationships between kinds of indexes

So far we have spoken of indexes for topic, place, time, and persons as if the indexes for these facets were separate and independent, but in practice they are not, except in primitive examples. In a mature topical index such as the Library of Congress Subject Headings system, the topic heading will be commonly combined with geographical and chronological qualifiers, e.g. *Architecture – Japan – Meiji period, 1868-1912*. In other words, subject headings may have geographical and temporal components as well as topical.

A place name gazetteer ordinarily indicates the kind of place (geographic "feature type") it is: castle, church, city, lake, city, etc. A physical feature is not the same as a topic, but any kind of feature can be treated as a topic. An individual castle is an instance of the category *castles*. Documents about castles generally may be helpful as well as any documents concerning this particular castle. And a discussion of the topic *castles* can be enriched by moving from the subject heading to the geographical feature type codes in the gazetteer in order to identify and to locate *instances* of castles in any region, so a mapping between feature types and subject headings can be useful. Since a well-designed gazetter will also have an indication of *when* that name was in use, entries in gazetteers, like subject headings, can have temporal and topical as well and geographical aspects.

The time period directory, which we modelled on gazetteer designs, has a coding for kind of event or period. So, as with gazetteer entries, a specific event (e.g. an earthquake) can be linked to subject headings both by proper name (e.g. *Lisbon Earthquake 1755*) and also the literature on that class of events (e.g. *Earthquakes*). Events are specific to geographic areas and so a proper time period dictionary will have geographical codings and it should be possible to link each event to both geographic subject headings and to gazetteer entries.

The texts of entries in biographical dictionaries are very rich in mentions of (i) kinds of activities, which could be linked to subject headings for that kind of activity; (ii) to places that could be linked to gazetteer entries and to geographic subject headings; (iii) to periods of time that could be linked to other, contemporaneous events via time period directories, timelines, and chronologies; and (iv) other people with whom the biographee interacted and for which biographical information could be found in biographical dictionaries and encyclopedias.

Subject indexes, place name gazetteers, time period directories, and biographical dictionaries are quite different genres for quite different aspects of reality, but we find geographical connections, chronological links, and topical affinities across all four. There is a large and useful agenda in finding ways to build effective infrastructure of

connections between these genres, because understanding requires a knowledge of context.

Conclusion and Agenda

The initial first role of metadata is as description, but descriptions can be manipulated to provide support for search and selection. A central agenda of information science has been the creation of descriptions and of indexes. Now, with a network environment there is new opportunity to extend this agenda into linking not only between between different indexes of the same kind but also between indexes of different kinds. We need to develop "best practices" and standards to link entries in thesauri, place-name gazetteers, time period directories, biographical dictionaries, and, especially, between these different genres. We need to build a better metadata infrastructure.

Acknowledgment

This essays draws heavily on a series of studies at the Electronic Cultural Atlas Initiative (Buckland & Lancaster 2004, 2006) under the leadership of Michael Buckland, Fredric C. Gey, and Ray R. Larson, partially supported by the U.S. federal Institute for Museum and Library Services: "Seamless Searching of Numeric and Textual Resources" (NLG 178); "Going Places in the Catalog: Improved Geographic Access" (LG-02-02-0035-02), and " Support for the Learner: What , Where, When, and Who" (LG-02-04-0041-04).

References

Bringing Lives to Light: Biography in Context. 2006. http://ecai.org/imls2006

Buckland, Michael, Aitao Chen, Fredric C. Gey & Ray R. Larson. Forthcoming. Search Across Different Media: Numeric Data Sets and Text Files. *Information Technology and Libraries* http://www.lita.org/ala/lita/litapublications/ital/italinformation.htm

Buckland, Michael, Aitao Chen, Fredric C. Gey, Ray R. Larson, Ruth Mostern & Vivien Petras. Forthcoming. Geographic Search: Catalogs, Gazetteers, and Maps. *College & Research Libraries* http://www.ala.org/ala/acrl/acrlpubs/crljournal/collegeresearch.htm

Buckland, Michael, Fredric C. Gey & Ray R. Larson. 2002. *Seamless Searching of Numeric and Textual Resources: Final Report on Institute of Museum and Library Services National Leadership Grant No. 178*. (Berkeley, CA: University of California, School of Information Management and Systems, 2002). http://metadata.sims.berkeley.edu/papers/SeamlessSearchFinalReport.pdf Accessed 4 November 2006.

Buckland, Michael, Fredric C, Gey, and Ray R. Larson. 2004. *Going Places in the Catalog: Improved Geographic Access: Final Report.* http://ecai.org/imls2002/imls2002-final_report.pdf Accessed 4 November 2006.

Buckland, Michael and Lewis Lancaster. 2004. Combining time, place, and topic: The Electronic Cultural Atlas Initiative. *D-Lib Magazine* 10, no. 5 (May 2004) http://www.dlib.org/dlib/may04/buckland/05buckland.html Accessed 2 November 2006.

Buckland, Michael & Lewis R. Lancaster. 2006. Advances in discovery: The Electronic Cultural Atlas Initiative experience. *First Monday* (August 2006) http://www.firstmonday.org/issues/issue11_8/buckland/index.html Accessed 2 November 2006.

Caplan, Priscilla. 2003. *Metadata Fundamentals for all Librarians*. Chicago: American Library Association.

Day, Ron. 2006. *Notes on Infrastructure and Development*. http://ella.slis.indiana.edu/~roday/infrastructure.html Accessed 2 November 2006.

Foucault, Michel. 1970. *The Order of Things: an Archaeology of the Human Sciences*. New York: Pantheon Books.

Haynes, David. 2004. *Metadata for Information Management and Retrieval*. London: Facet Publishing.

National Library of Medicine. 2006. *Unified Medical Language System.* Factsheet. [Webpage]. http://www.nlm.nih.gov/pubs/factsheets/umls.html Accessed 29 October 2006.

National Science Foundation. 2003. *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. ("Atkins Report"). http://www.nsf.gov/cise/sci/reports/atkins.pdf Accessed 2 November 2006.

Petras, Vivien, Ray Larson, & Michael Buckland. 2006. Time Period Directories: A Metadata Infrastructure for Placing Events in Temporal and Geographic Context. In: *Opening Information Horizons: Joint Conference on Digital Libraries (JCDL), Chapel Hill, NC, June 11-15, 2006.* http://metadata.sims.berkeley.edu/tpdJCDL06.pdf Accessed 4 November 2006.

Text Encoding Initiative Consortium. 2006. *Report on XML Mark-up of Biographical and Prosopographical Data*. http://www.tei-c.org/Activities/PERS/persw02.xml Accessed 2 November 2006.

Zerneke, Jeannette L., Michael K. Buckland & Kim Carl. 2006. Temporally Dynamic Maps: The Electronic Cultural Atlas Initiative Experience. *Human IT* 8.3 (2006): 83–94. http://www.hb.se/bhs/ith/3-8/jzmbkc.pdf Accessed 4 November 2006.