IJDC | *General Article*

# Controlled Vocabulary Standards for Anthropological Datasets

Celia Emmelhainz
Kent State University

## Abstract

This article seeks to outline the use of controlled vocabulary standards for qualitative datasets in cultural anthropology, which are increasingly held in researcher-accessible government repositories and online digital libraries. As a humanistic science that can address almost any aspect of life with meaning to humans, cultural anthropology has proven difficult for librarians and archivists to effectively organize. Yet as anthropology moves onto the web, the challenge of organizing and curating information within the field only grows. In considering the subject classification of digital information in anthropology, I ask how we might best use controlled vocabularies for indexing digital anthropological data. After a brief discussion of likely concerns, I outline thesauri which may potentially be used for vocabulary control in metadata fields for language, location, culture, researcher, and subject. The article concludes with recommendations for those existing thesauri most suitable to provide a controlled vocabulary for describing digital objects in the anthropological world.

*Received* 02 December 2013 | *Accepted* 26 March 2014

Correspondence should be addressed to Celia Emmelhainz, 4, 43-a Str., Astana 010000 Kazakhstan. Email: cemmelha@kent.edu or celia.emmelhainz@gmail.com

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/



 doi:10.2218/ijdc.v9i1.290

# Introduction

In this article I consider current digital archiving practices in anthropology, a field of academic study which developed in the 19th century in order to compare world cultures. Subfields include the study of human cultures, bodies, languages and archaeology (Ingold, 1994), although this paper addresses itself to the subfield of cultural anthropology. As a humanistic science that can address any aspect of human life, cultural anthropology has proven difficult to organize effectively (Kotter, 2002). As a once paper-based discipline moves onto the web, the challenge of organizing information within anthropology grows even more complex. In examining the subject classification of digital information in anthropology, this paper asks how we may best provide controlled vocabularies for subject indexing of digital anthropological data. After outlining current organizational issues in the field, I link available thesauri to the emerging metadata fields in anthropological datasets which are most in need of a controlled vocabulary standard.

# Archiving Anthropological Data

The major task of a cultural anthropologist is first to develop questions about the world, then to seek answers about how communities address these concerns in diverse environments, and then to write about this fieldwork in a theoretically compelling way. Fieldwork is here defined as a period of time during which anthropologists seek cultural immersion in their target setting, for research purposes. A massive set of resulting document include fieldnotes (a diary and analysis of daily interactions), interview notes, photos, videos, GIS data and other ephemera. A single field project may result in a print or electronic dataset comprising thousands of discrete items on a range of topics, including 'audio-recordings of interviews, written transcripts of those interviews and, in some cases, annotations or codings already undertaken' (Cheshire, 2009).

Historically, the field data of notable anthropologists have been gathered in physical archives, which may be hard to access and prone to decay (Zeitlyn, 2012); the papers of minor field researchers are even more easily lost. As anthropologists now collect most raw data directly in digital form (with daily notes on tablets and voice recorders, and pictures on cameras and iPhones) there is an increased danger of data loss. Such data loss often occurs due to file type incompatibility, misplacement of inherited files, or hardware failure (Cliggett, 2013).

In the past twenty years, some researchers have begun to gather qualitative datasets in long-term digital archives, with major collections hosted by the UK and Australian governments, as well as by major US research institutes, such as Harvard's Murray Center and Michigan's Inter-university Consortium for Political and Social Research (ICPSR). With a nod to established standards, such as Dublin Core, a range of metadata schemas have been established for qualitative data archiving, as in the following extensive list of descriptive elements recommended by ICPSR for archiving with their institution, the University of Michigan:

> Names of principal investigators; project title; funding sources; data collector; project description; sampling procedure; variables and weighting;

> date, geographic location, and time; links to other data sources; units of analysis (and subject); links to related publications; technical metadata; links to data collection instruments; flowchart of data collection; link to list of abbreviations and coding instrument (summarized from ICPSR, 2012).

One anthropologist, Lisa Cliggett (2011), has further recommended adding an element for kinship grouping within datasets that concern kin-based or agrarian people groups. In describing anthropological records, Wade Kotter (2002) recommends a set of data fields that include "geographical location, ethnic group, time period, methodological approach and theoretical perspective," asserting that standard vocabularies may be needed for any of these fields.

The current leader in digitally archiving and sharing qualitative datasets is QualiData[1], a government-funded British project that has promoted the QuDEx[2] metadata schema for qualitative archiving. This schema provides for a top-element <codeCollection>, comprising multiple <code> elements, and defines a <code> as:

> 'A short alphanumeric string, usually a single word [which] may be assigned to a segment or document ... A code may optionally be taken from a controlled vocabulary defined under @ authority' (Corti, 2008).

However, although the major qualitative dataset archiving centres at QualiData (UK), ICPSR (USA), and AQuA[3] (Australia) each affirm that authority control should be used, I find no evident links or recommendations from these archives as to which standards could be adopted for a controlled vocabulary; the choice of such standards seems to be left to the non-technical user. The remainder of this paper outlines concerns with available controlled vocabularies and then assesses the thesauri that might be used in controlling metadata elements for digital archives of anthropological field data.

## Thesauri for Controlled Vocabularies

Controlled vocabularies can be defined as lists of preferred terms which are used to populate common metadata elements, including subject or research type; these make subject searches more effective (Taylor and Joudrey, 2009). Subjects may comprise topics as varied as concepts, names, location, chronological elements, and form; in anthropological datasets, each of these elements could benefit from controlled vocabularies.

However, a major challenge in setting vocabulary standards for anthropology is the field's strong link to the interpretivist paradigm of research, which 'emphasizes human subjectivity and the meanings people attached to the world' (Cheshire, 2009). If qualitative researchers seek not to define clear constructs of knowledge, but instead to explore the social tags by which people give meaning to their world, why would one assign a single controlled vocabulary to datasets which, in effect, simply expose the many non-overlapping ontologies by which people live?

Such issues with the interpretivist paradigm are likely to remain an ongoing concern for many practicing anthropologists, but librarians and archivists may find strong benefits to adopting a clear set of controlled vocabulary standards when describing these

1 QualiData: http://www.esds.ac.uk/qualidata/about/introduction.asp
2 QuDEx: http://www.data-archive.ac.uk/create-manage/projects/qudex
3 AquA: http://researchdata.ands.org.au/australian-social-science-data-archive-network-extension-and-sub-archive-development

professionals' datasets – and even the most critical anthropology faculty and students will ultimately benefit from being able to search databases for past research using a clearly-defined set of terms. While subject terms and concerns within the field will continue to change over time, I would argue that defining vocabularies is still a critical exercise in order to effectively provide researchers with subject access to relevant datasets.

Yet even if the necessity of a controlled vocabulary is acknowledged by researchers, the complexity of inter-related descriptive metadata elements may remain a concern. Even a book's title can hint at 'a discipline, a subtopic, several geographic locations, and a comparative relationship' (Taylor and Joudrey, 2009), and such a problem is surely compounded when classifying complex sets or subsets of research products for long-term use. Kotter (2002) believes that the complex subject material of anthropology results in a need for post-coordinated depth indexing, proposing a faceted classification system that could effectively provide access to documents such as:

> '[an] article on labor allocation focuses on the Lancang region of Southwest China; refers to an ethnic group known as the Qhawqhat Lahu; is concerned with the present day; utilizes data gathered through participant observation ... and adopts the interpretive framework sometimes referred to as critical theory' (Kotter, 2002).

Yet while he lists many possible aspects of his personal faceted classification scheme, Kotter's plan has not yet been implemented or published. The reader is left to consider other thesauri that could more easily organize anthropological datasets.

## Recommendations on Specific Vocabulary Standards

To address this issue, I highlight five key metadata elements for fieldwork datasets that could benefit from adopting a standard of vocabulary control. These are language, location, culture, researcher, and subject. A brief discussion of each element reviews possible thesauri, and notes the strengths and weaknesses of each. No single option is ideal, but a final table highlights the recommended options that may provide an immediate way of structuring access to data, at least until more subject-specific vocabularies are developed.

### Language

Kotter (2002) recommends using the Compendium of the World's Languages[4] as a basis for developing an anthropological controlled vocabulary, but the Compendium appears to be expensive, unavailable in database format, and focused on describing major languages rather than on developing a widely standardized codification system. For this reason, the widely standardized ISO 639 codes are recommended.

### Location

The Getty Thesaurus of Geographic Names[5] has a hierarchical classification which identifies larger cities, such as the city of Ekibastuz in Kazakhstan, but misses many smaller locations; it may be helpful for classifying at the regional level.

4 Compendium of World Languages: https://archive.org/details/rosettaproject_bla_morsyn-1
5 Getting Thesaurus of Geographic Names: http://www.getty.edu/vow/TGNSearchPage.jsp

**Culture**

Anthropologists have historically been concerned with sorting human behaviour via "culture," and an Outline of World Cultures[6] (OWC) was created by George Murdock in 1954 to number and organize world cultures into distinct ethnic groups by country-based location[7]. This system attempts to account for how cultural groups diverge in practice given their location and identity ties (cf. Cunnar, 2014), but one major disadvantage is that the OWC numbering system is not searchable or indexed freely online.[8]

**Researcher**

There is currently no name authority file that systematically lists researchers of culture. The Library of Congress Name Authority Headings[9] do include anthropologists with published monographs or dissertations, but may not include those with large datasets who only published in journal articles. These Name Authority Headings may serve as a guideline, but a disciplinary extension of this namespace would better account for all researchers in the field.

**Subject**

The central decision to be made in subject classification is whether depth or breadth is most important, given user needs. The Library of Congress Subject Headings[10] (LCSH) is strongest for breadth of general use and would be suitable as a vocabulary for cross-disciplinary datasets, but lacks the organized focus on anthropological concerns typical of a more specialized thesaurus.

For more specific terms, Kotter (2002) recommends the Thematic List of Descriptors – Anthropology (UNESCO, 1989). However, this $700 volume is only in print, and has not been updated in fifteen years. More recently, ProQuest has combined the above thematic list with three other UNESCO lists in economics, political science, and sociology, forming their subject headings for the International Bibliography of the Social Sciences (IBSS) database[11]. IBSS advertises a combined 10,000 linked subject terms across the fields of anthropology, economics, political science and sociology. However, this social sciences database is subscription-only (with ProQuest) and the subject list itself does not seem to be publicly accessible for browsing or adaptation. For this reason, neither UNESCO's thematic list nor the IBSS subject list can currently be recommended for use in classifying anthropological datasets.

In folklore, the American Folklore Society's Ethnographic Thesaurus[12] provides one acceptable classification system for cultural and ethnographic subjects, with a focus on traditional roles, rites, and places. It remains a valid option, but is strongest only in the folklore subfield of the discipline.

6 Outline of World Cultures: http://catalog.hathitrust.org/Record/009055937

7 For example, see implementation at: http://hraf.yale.edu/

8 The Joshua Project is the other classifying scheme which groups cultures by ethnicity and location in similar depth (see http://joshuaproject.net/). However, while easily filtered and freely available online, the project's overwhelming focus on Christian conversion of indigenous groups makes it inappropriate for use with most student or researcher populations.

9 Library of Congress Name Authority Headings: http://authorities.loc.gov/

10 Library of Congress Subject Headings: http://id.loc.gov/authorities/subjects.html

11 International Bibliography of the Social Sciences: http://www.proquest.com/products-services/ibss-set-c.html

12 Ethnographic Thesaurus: http://www.openfolklore.org/et/tree.htm

And finally, Yale's eHRAF ethnographic collection uses the extremely detailed Outline of Cultural Materials[13] to sort 700 categories of "cultural" items, systems, and artifacts in anthropological research (cf. Roe, 2007). A printed list as well as a detailed tree-system[14] of classification standards online allows for detailed indexing of common anthropological topics, and is recommended as a reasonable set of subject terms in indexing anthropological datasets.

**Table 1.** Summary Table of Recommended Controlled Vocabulary Standards.

| Metadata element | Recommended standard and its source, if any |
|---|---|
| <language> | Use the ISO 639 standard. |
| <location> | Use the Thesaurus of Geographic Names (TGN). |
| <culture> | Use the Outline of World Cultures (OWC). |
| <researcher> | Use Library of Congress Namespace (LCCN) for published authors; create a Name Authority File for other disciplinary researchers. |
| <subject> | Use Library of Congress Subject Headings (LCSH) for general sets. |
| <subject> | Use AFS's Ethnographic Thesaurus for folklore datasets. |
| <subject> | Use the Outline of Cultural Materials (OCM) for anthropological sets. |

As evident above, there are multiple options for vocabulary control in anthropological datasets – and given the diversity of the field, the development of a single controlled vocabulary standard or classification scheme is unlikely to occur. However, this discussion of possible standards is intended to assist researchers in adopting the most appropriate existing thesauri for their needs, and also intended to encourage the future development of vocabularies more closely adapted to the highly specified classification needs of anthropological researchers.

## Conclusion

This brief discussion is intended to direct the publisher of qualitative datasets and digitally-curated fieldwork data to the most appropriate existing controlled vocabulary standards for use in structuring anthropological metadata. While "controlled vocabulary" is frequently recommended to those archiving qualitative data, a recent and published discussion of existing standards seems to be absent from the literature. Given the breadth of anthropological research topics, it is recommended to use the Outline of Cultural Materials or Library of Congress Subject Headings for primary subject headings, and then to supplement with sub-field thesauri, such as the Ethnographic Thesaurus for folklore, if warranted. For one example, see the AustKin database developed to define the multiplicity of aboriginal kinship terms (Dousett et al., 2010). In addition, there is no name authority file that currently addresses long-term anthropological researchers who have not produced monographs, and one should be considered. An online registry of anthropological thesauri at one of the major research centres could further promote the use of controlled vocabularies for anthropological

13 Outline of Cultural Materials: http://www.durhamtech.edu/dtcclibrary/ehraflist.pdf
14 See: http://hraf.yale.edu/online-databases/ehraf-world-cultures/outline-of-cultural-materials/#id197

data. Each of these steps will provide researchers and archivists with more fine-grained control and discoverability for both public and privately-held digital fieldwork datasets.

## Acknowledgements

Many thanks to Lisa Cliggett and Frank Lambert for discussion on these issues.

## References

Cheshire, L. (2009). *Archiving qualitative data: Prospects and challenges of data preservation and sharing among Australian qualitative researchers*. Retrieved from Australian Social Science Data Archive website: http://www.assda.edu.au/forms/AQuAQualitativeArchiving_DiscussionPaper_FinalNov09.pdf

Cliggett, L. (2011). *Strategies of data archiving for cultural anthropology: Using Gwembe Tonga research project (GTRP) data.* Project proposal submitted to the National Science Foundation.

Cliggett, L. (2013). Qualitative data archiving in the digital age: Strategies for data preservation and sharing. *The Qualitative Report*, *18*(24), How To Article 1. Retrieved from http://www.nova.edu/ssss/QR/QR18/cliggett1.pdf

Corti, L. (2008). *Data Exchange Tools and Utilities (DeXT): Final Report.* Retrieved from UK Data Archive website: http://data-archive.ac.uk/media/1681/DexT_finalreport_JISC.pdf

Cunnar, C. (2014). Community life in Tajikistan, Uzbekistan, Kazakhstan, Kyrgyzstan, and Turkmenistan [Web log post]. Retrieved from Human Relations Area Files blog: http://hraf.yale.edu/community-life-in-tajikistan-uzbekistan-kazakhstan-kyrgyzstan-and-turkmenistan/

Dousett, L., Ehndery, R., Bowern, C., Koch, H., & McConvell, P. (2010). Developing a database for Australian indigenous kinship terminology: The AustKin project. *Australian Aboriginal Studies, 1,* 42–56. Retrieved from http://search.informit.com.au/documentSummary;dn=337481357042539;res=IELIND

Inter-university Consortium for Political and Social Research. (2012). *Guide to social science data preparation and archiving: Best practice throughout the data life cycle* (5th ed.). Ann Arbor, MI: Author. Retrieved from http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf

Ingold, T. (1994). *Companion encyclopedia of anthropology*. London and New York: Routledge.

Kotter, W. (2002). Improving subject access in anthropology. *Behavioral & Social Sciences Librarian*, *20*(2), 1–15. doi:10.1300/J103v20n02_01

Murdock, G.P. (1954). *Outline of world cultures*. New Haven, CN: Human Relations Area Files.

Roe, S.K. (2007). A brief history of an ethnographic database. *Behavioral & Social Sciences Librarian, 25*(2), 47–77. doi:10.1300/J103v25n02_03

Taylor, A. & Joudrey, D. (2009). *The organization of information* (3rd ed.). Westport, CN: Libraries Unlimited.

UNESCO. (1989). *Thematic list of descriptors, anthropology*. London: Routledge.

Zeitlyn, D. (2012). Anthropology in and of the archives: Possible futures and contingent pasts. Archives as anthropological surrogates. *Annual Review of Anthropology, 41,* 461–480. doi:10.1146/annurev-anthro-092611-145721