Journal of World-Historical Information
JWHI

RESEARCH REPORT

Tonia Sutherland

Complexities in Creating and Defining Topical Metadata:

Practices for a World-Historical Data Resource

The Collaborative for Historical Information and Analysis (CHIA), a large-scale digital humanities project, aims to link world-historical data in the social sciences, natural sciences, and humanities; allow researchers to draw new connections and new conclusions from analyzing large-scale aggregated datasets; and provide for the long-term preservation of historical data. To accomplish these tasks, CHIA requires a coherent metadata framework to link data to their sources and each other (CHIA, NSF Grant Proposal). Work on geospatial gazetteers and temporal ontologies has already begun; however, in order to ensure the longevity of CHIA data, allow users to understand and access the data, and publish datasets for future use, defining a controlled vocabulary of topical metadata for CHIA's data is vital. This research aims to define topical metadata practices for CHIA through the creation and definition of a controlled vocabulary that will add value to CHIA data and allow for future growth and change.

Volume 2-3, No. 1 (2014-2015) | ISSN 2169-0812 (online)
DOI 10.5195/jwhi.2015.22 | http://jwhi.pitt.edu

The Collaborative for Historical Information and Analysis

The Collaborative for Historical Information and Analysis (CHIA) is effectively a research and development project joint-funded by the National Science Foundation (2010) and the University of Pittsburgh School of Arts and Sciences. CHIA's Human Systems Data Resource, currently under development, will use both crowdsourcing tools and fetched data for large-scale data aggregation and analysis. This system offers researchers the potential to edit and link datasets for new kinds of data analysis as well as encouraging collaborative research through open source toolboxes and crowdsourcing. Ultimately, CHIA aims to bring together disparate data from around the world from 1500-present. To accomplish this, CHIA currently brings together user-contributed data from the social sciences, natural sciences, and humanities. Current contributors include historians, geographers, political scientists, economists, and information scholars and scientists.

Functioning in part as a data archive, CHIA relies on the Dataverse Network project initiated by one of its partner institutions, the Institute for Quantitative Social Sciences at Harvard University, for long-term preservation of world-historical data. The datasets in CHIA's World-Historical Dataverse are diverse: they currently comprise data on population (including age and gender), migration, mortality, literacy, religion, disease, opium production and consumption, taxes, wages, trade, climate, railways, hospital beds, drug shops, correlates of war, and world development indicators. Researchers who deposit data into CHIA's World-Historical Dataverse are bound by existing internal documentation policies and procedures which provide basic descriptive metadata for submitted datasets.[1]

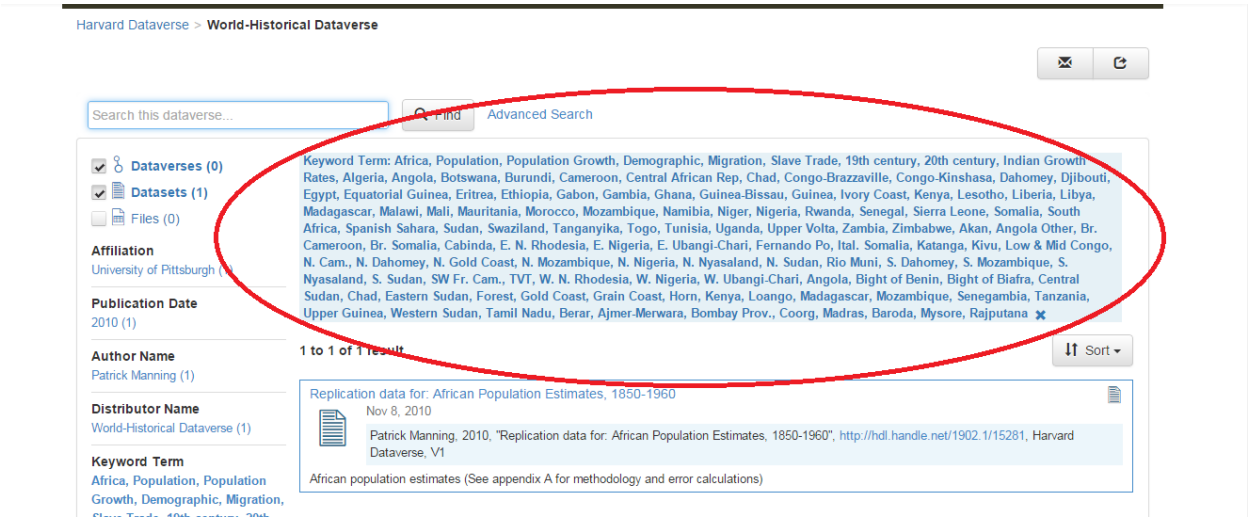

Figure 1. Current display of search results by "keyword term" in the CHIA instance of Harvard's Dataverse. These terms, represented as a single hyperlink, are the actual column headers for CHIA's African Population Estimates (1850-1960) dataset.

Some of CHIA's main goals and challenges, from a data curation standpoint, are: simultaneously sustaining the documentation of and increasing the quality and dependability of the datasets being contributed to the project; building robust, comprehensive tools to address the heterogeneity of data and the complexity of merging data from multiple contributors; and reconciling the idea of creating a centralized data resource at a time when computing and storage is increasingly distributed. Meeting these goals has required assembling and managing a team of developers and nurturing a worldwide community of scholars for data gathering and analysis.

Problem Statement – Creating and Defining a Topical Ontology

Metadata is used for the discovery, identification, and representation of data in a given system or structure. Within the broader category of descriptive metadata, subject—or topical—metadata generally refers to controlled vocabularies that describe and define the subject or topic of the data. There is no existing controlled vocabulary or metadata schema that is both global and historical, and certainly not one that addresses CHIA's global timeline comprising the years 1500-1950. As an alternative to controlled vocabularies, user-generated folksonomies and social tags have emerged as possible solutions to the limitations of existing vocabularies. One clear benefit of folksonomies, particularly when compared to controlled vocabularies such as the Library of Congress Subject Headings, is their capacity to allow users to name their own resources in their own terms. However, finding subject terms for classificatory metadata from user-generated or crowd-sourced tags also presents challenges. Because social tags are not generated from controlled vocabularies, there are issues—such as internal consistency and interoperability—that must be addressed before it is possible to identify the best terms to represent the content of a data resource. The ultimate success of a project such as CHIA relies heavily on descriptive tools that are not only historically accurate, culturally sensitive, and diverse, but tools that are also attuned to the unique challenges of electronic resources: as a large-scale aggregated and heterogeneous data resource, the bulk of CHIA's content is either born-digital or digitized.

The Association for Library Collections and Technical Services Subject Access Committee (ALCTS-SAC) Subcommittee on Metadata and Subject Analysis recommends that subject metadata should be both straightforward and intuitive. On one hand, Library of Congress Subject Headings (LCSH) have met the ALCTS-SAC criteria. They offer a rich vocabulary with pre-coordinated terms, synonyms, and a syndetic structure that allows for cross-referencing. LCSH improve discovery and can be applied in many parts of the world because LCSH has been translated in several languages. LCSH's complex syntax is, however, ill-suited to an Internet environment (ALCTS-SAC). Moreover, LCSH critics such as Sanford Berman and Hope Olson have claimed that LCSH reflect and serve a mainstream audience, lacking terms for, or misrepresenting, groups on the margins and failing to account for cultural change over time.[2] For example, the terms "Darfur Genocide, 2003-" (to replace "Sudan--History--Darfur Conflict, 2003-") and "Industrial Pollution" (no current term) were submitted as subject heading proposals by Berman, a self-described "radical librarian" and author of Prejudices and Antipathies: A Tract on the LC Subject Heads Concerning People to the Library of Congress in the years 2007 and 2008, respectively.[3] To date, neither term has been authorized as an LCSH.

This research seeks to construct a practice for an ongoing controlled vocabulary that continually supports CHIA's work and future development. To accomplish this task, the research will 1) answer questions about the global and historical categorizations that have resulted in existing controlled vocabularies (classification practices have a long and complex history); 2) analyze the implications of these categorical vocabularies for a world-historical data project; 3) uncover the requirements for topical metadata for a world-historical data resource, moving the CHIA project closer to a comprehensive metadata framework; and 4) make suggestions about the way forward, creating and defining standards for CHIA data to include more human and cultural variety than existing controlled vocabularies.

From a theoretical perspective, this study is innovative in that it analyzes subject categorization in several different thesauri of controlled vocabularies, exploring the ways in which these vocabularies produce a particular kind of hegemonic discourse, and considering whether multiple voices are able to emerge within those structures. In contrast to the unimodal descriptive language that tends to comprise existing controlled vocabularies, CHIA's data requires a metadata framework that allows multiple voices to be heard, at different scales, simultaneously. While

each CHIA dataset represents at least one historical voice, together, the CHIA data—as a world-historical data resource—will ultimately encompass a polyphony of voices. For researchers to engage with any individual voice within this collection of CHIA data, appropriately grounded metadata is imperative.

Methodology

One approach to solving the problem of a topical ontology for CHIA is to examine existing vocabularies qualitatively, taking into consideration who constructs vocabularies and their categories and who ultimately authorizes these descriptive constructs. Another approach is to consider existing vocabularies collectively and comparatively, locating areas of connection and contention, and uncovering areas where these descriptive tools allow for the multiple perspectives represented in the CHIA data to be used and analyzed. The research relies on controlled vocabularies such as LCSH and the UNESCO Thesaurus, and may particularly consider derivations of the UNESCO Thesaurus such as the National Digital Archive of Datasets Thesaurus (NDAD). Other thesauri of controlled vocabularies such as HASSET, the Humanities and Social Science Electronic Thesaurus, will also be considered. Selected terms across several thesauri will serve as data points for comparative analysis. Findings will satisfy the practical needs of the CHIA project, add value to the CHIA data, and offer new kinds of uses for CHIA data that are only possible with a topical controlled vocabulary that allows for evolution and growth.

Existing Databases and Classification Schemas

Library of Congress Subject Headings (LCSH). The Library of Congress Subject Headings (LCSH) database has been actively maintained since 1898 (first in print form, now electronically), with its primary purpose being the description and cataloging of materials held at the Library of Congress. LCSH is also used by other libraries, both in the United States and internationally, to provide subject access to their collections.[4] LCSH includes personal and corporate name authority entries, geospatial entries (which are updated on an as-needed basis), and subject/topical entries. Authorized terms are organized hierarchically in LSCH with broader terms (BT), narrower terms (NT), and related terms (RT) associated with many—but not all—entries.

Authority entries are constructed according to specific practices: subject headings may consist of one word or several with a one-word heading typically being a noun; concepts are named in the singular and objects in the plural; two-word headings typically comprise an adjective and a noun which may appear in typical word order (*Local ordinance, Molecular biology*) or in inverted form. Inverted forms are most commonly used for languages, nationalities, or other ethnic adjectives such as *Art, French* or *Lullabies, Urdu*. Inversions are also used to qualify time periods such as *Painting, Renaissance*. LoC's original organizational scheme would have reflected dictionary order rather than the current alphabet-class order. Instead, the classification order reflects a reluctance on the part of LoC to separate related entries. As such, many headings were originally constructed in a manner that placed the classification first, *Photography—Studios* and *Railroads—Timetables* are two examples. [5]

The current edition of Library of Congress Subject Headings (LCSH 37) is the 37th edition of the authority file and contains headings established by LoC through January 2015. The headings were obtained by creating a file consisting of all subject heading and subdivision records (those which have been verified only) in the subject authority file at LoC (approximately 337,354 authority records). The database from which the headings in the 37th edition were drawn indicates that the file contains approximately 24,018 personal name headings of which 22,854 represent family names, 9,454 corporate headings, 9 meeting or conference headings, 485 uniform titles, 239, 916 topical subject headings, and 60,354 geographic subject headings. Additionally, there are 4,360 general "see also" references.[6]

LCSH is considered "the most widely adopted subject indexing language in the world;" proposals for additions and changes are routinely reviewed in the Policy and Standards Division of the Library of Congress and lists of new and changed subject headings are posted on the Library of Congress (LoC) Cataloging and Acquisitions web site as they are approved. [7] Proposals for changes to existing headings may be submitted by cooperating libraries and the general publics. The learning curve for those wishing to make non-specialist suggestions for change is quite high, demanding a depth and breadth of knowledge relating to bibliographic cataloging practices that even many practicing librarians do not maintain. For example, the April 2015 Tentative Monthly List includes the following suggested changes:

LCSH Authority Entry	**Proposed Change, April 2015**
Acala (Buddhist deity)	CANCEL HEADING
BT South America--Antiquities	ADD FIELD
BT Ethnology--Indonesia	
Patasiwa Alfoeren language	ADD FIELD
Sapalewa language	ADD FIELD

Table 1. From Library of Congress Subject Headings

The first example suggests cancelling an existing subject heading about the Buddhist deity Acala while the second example calls for adding a broad term "South America--Antiquities" under which narrower terms might fall. In the third example, the suggested change to the broad term "Ethnology--Indonesia" is to add narrower terms relating to two Indonesian languages: Patasiwa Alfoeren and Sapalewa.

United Nations Educational, Scientific, and Cultural Organization (UNESCO) Thesaurus. Per UNESCO, the UNESCO Thesaurus is a "controlled and structured list of terms used in subject analysis and retrieval of documents and publications in the fields of education, culture, natural sciences, social and human sciences, communication and information." This multidisciplinary thesaurus is updated on an ongoing basis and reflects the evolution of UNESCO's programs and activities. The UNESCO Thesaurus contains 7,000 terms in English and in Russian, as well as 8,600 terms in French and in Spanish.[8]

The entire UNESCO thesaurus is available for download.[9] Additionally, thesaurus entries can be viewed as an alphabetical list, a permuted list, or as a hierarchical arrangement. The permuted list displays descriptors and non-descriptors alphabetically by each significant word. For example, the compound term "intangible cultural heritage" can be located via each individual term contained within it. Similarly, a search on "cultural" will list automatically all the compound terms containing this word (i.e. cultural heritage, movable cultural property, etc.).

The hierarchical arrangement displays UNESCO's seven major domains—education; science; culture; social and human sciences; information and communication; politics, law and economics; and countries and country groupings—subdivided into microthesauri (MT) each containing variable numbers of top-level terms—those descriptors without associated broader terms (BT). Each top-level term is associated with appropriate "used for" terms, if any, followed by a descending hierarchy of descriptors, each preceded by an NT—"narrower term"—distinction. Finally, scope notes (SN), if available, provide guidelines for use.[10] For example, the BT "academic achievement" has the following hierarchical entry:

MT 1.65 Educational evaluation
Academic Achievement
UF Academic performance
UF Student achievement
UF Student progress
NT1 Academic failure
NT2 Dropping out
NT2 Grade repetition

Table 2. From UK Data Archive: Humanities and Social Science Electronic Thesaurus (HASSET)

The Humanities and Social Science Electronic Thesaurus (HASSET) is a subject thesaurus developed by the UK Data Archive at the University of Essex over the past 20 years. The thesaurus' subject coverage reflects the subject content of the UK Data Service holdings focusing on the social sciences and humanities. Coverage is most comprehensive in the "core subject areas of social science disciplines: politics, sociology, economics, education, law, crime, demography, health, employment, and, increasingly, technology and environment."[11] Humanities disciplines such as history and linguistics also have a prominent presence. HASSET uses standard hierarchical relationships: TT (top term); BT (broader term); NT (narrower term); RT (related term); USE (from non-preferred term to preferred term); UF (from preferred term to non-preferred term).[12] While new HASSET terms are developed regularly in order to cover the ever-growing collection of the UK Data Archive, HASSET is also a licensed product that must be purchased for ongoing access and use. HASSET employs the Simple Knowledge Organization System (SKOS). SKOS is a language designed to represent the structure and content of thesauri and other classification resources in a standardized way. It uses Resource Description Framework (RDF) to make such resources comparable, and to facilitate interaction. Both SKOS and RDF are made up of triples, each of which includes a

Subject	Predicate	Object (Literal or URL)
http://lod.data-archive.ack.uk/skoshasset/9c7bbfc3-325f-49eb-bfcd-078232ea	rdf:type	skos:Concept
http://lod.data-archive.ack.uk/skoshasset/e5722ba3-f36b-43e7-911b-a79c83977fe3	rdf:type	skos:Concept
http://lod.data-archive.ack.uk/skoshasset/9c7bbfc3-325f-49eb-bfcd-078232ea	skos:prefLabel	HEALTH
http://lod.data-archive.ack.uk/skoshasset/e5722ba3-f36b-43e7-911b-a79c83977fe3	skos:prefLabel	ORAL HEALTH
http://lod.data-archive.ack.uk/skoshasset/9c7bbfc3-325f-49eb-bfcd-078232ea	skos:narrower	http://lod.data-archive.ack.uk /skoshasset/e5722ba3-f36b-43e7-911b-a79c83977fe3

Table 3. Example of encoding in SKOS

subject, predicate and an object. For example, the relationship of ORAL HEALTH as a Narrower Term of HEALTH would be encoded in SKOS as shown in Table 3.

The Virtual International Authority File (VIAF). The Virtual International Authority File (VIAF) is a collaborative effort between national libraries and organizations, each contributing name authority files. VIAF allows users to repurpose bibliographic data produced by libraries serving different language communities. In VIAF, contributed authority data for each entity is linked together into a "super" authority record. As of 2012, VIAF contributors numbered 34 agencies in 29 countries. VIAF creators—Library of Congress, Deutsche Nationalbibliothek (DNB), the Bibliothèque nationale de France (BNF) and OCLC—envisioned VIAF as a "building block for the Semantic Web to enable switching of the displayed form of names for persons to the preferred language and script of the Web user."[13]

Other Thesauri, Glossaries, Ontologies and Controlled Vocabulary Databases[14]

Thesauri

- UNESCO-IBE Education Thesaurus: important indexing tool in the education field
- Food and Agriculture Organization Thesaurus: AGROVOC - Multilingual thesaurus of FAO, the Food and Agriculture Organization of the United Nations in 14 languages.
- European Language Social Science Thesaurus (ELSST) is a broad-based multilingual thesaurus for the social sciences.
- General Multilingual Environmental Thesaurus: GEMET - General Multilingual Environmental Thesaurus, in 19 European languages, developed by the European Environment Agency (EEA) and the European Topic Centre on Catalogue of Data Sources (ETC/CDS), in collaboration with European Union member countries and the UNEP Infoterra.
- NASA Thesaurus: The scope of this controlled vocabulary includes aerospace engineering in addition to supporting areas of engineering and physics, the natural space sciences (astronomy, astrophysics, planetary science), Earth sciences, and the biological sciences. The NASA Thesaurus contains over 18,400 subject terms, 4,300 definitions, and more than 4,500 USE cross references.
- Population Multilingual Thesaurus: POPIN - Population Multilingual Thesaurus: UN Population Information Network (PDF file - English, French and Spanish)
- Refugee Thesaurus: ITRT - International Thesaurus of Refugee Terminology (English, French, Spanish)
- UK Archival Thesaurus: UKAT - controlled vocabulary which archives can use when indexing their collections and catalogues
- UN Bibliographic Information System Thesaurus: UNBIS - United Nations Bibliographic Information System Thesaurus (Arabic, Chinese, English, French, Russian and Spanish)
- World Bank Thesaurus: The World Bank Group's Thesaurus contains 87,000 terms covering 30 knowledge domains

Glossaries / Terminology database

- International Atomic Energy Agency Safety Glossary: The IAEA Safety Glossary lists terms commonly used in safety related publications - (English only)
- Terminology database of the International Labour Organization: ILOTERM- Terminology database of the International Labour Organization (English, French, German and Spanish)
- International Monetary Fund Terminology: Over 4,500 records of terms useful to translators working with International Monetary Fund (IMF) material (English, French, German, Portuguese and Spanish)
- United Nations glossaries: United Nations interpreters' resource page—bilingual and multilingual glossaries (Chinese, English, French, Russian and Spanish)
- United Nations Multilingual Terminology Database: UNTERM - United Nations Multilingual Terminology Database (Arabic, Chinese, English, French, Spanish and Russian)
- World Trade Organization Glossary: This glossary is designed to help understanding of some of the terms used in the WTO and in international trade

NOTES

[1] See: Dataverse, "World-Historical Dataverse," https://dataverse.harvard.edu/dataverse/worldhistorical.

[2] See also: Adlera, Melissa. 2009. "Transcending Library Catalogs: A Comparative Study of Controlled Terms in Library of Congress Subject Headings and User-Generated Tags in LibraryThing for Transgender Books." Journal of Web Librarianship 3(4): 309-331.

[3] Berman, Sanford. 2008. "Personal LCSH scorecard." http://jenna.openflows.com/files/lcshscorecard080415.pdf

[4] Library of Congress Linked Data Service, "Library of Congress Subject Headings," http://id.loc.gov/authorities/subjects.html. International uses of LCSH are often done in translation, effectively making LCSH at least partly available in several languages.

[5] Library of Congress, "Introduction to Library of Congress Subject Headings," http://www.loc.gov/aba/publications/FreeLCSH/lcshintro.pdf, viii.

[6] Library of Congress, "Introduction to Library of Congress Subject Headings," http://www.loc.gov/aba/publications/FreeLCSH/lcshintro.pdf, vii.

[7] Library of Congress, "Subject Headings & Genre/Form Terms," http://www.loc.gov/aba/cataloging/subject/.

[8] UNESCO, "UNESCO Thesaurus," http://databases.unesco.org/thesaurus/.

[9] Please see: Simple Knowledge Organization System (SKOS), "UNESCO Thesaurus," http://skos.um.es/unescothes/downloads.php. The RDF/XLM version is 5.1 MB in size. It is also available for download as a Turtle file.

[10] "Used For" is denoted by UF in the UNESCO Thesaurus. UNESCO, like LoC, also uses RT, or "Related Terms." The point of descriptive departure is in the synonymic implications of UF in contrast to terms that are conceptually related, but not necessarily synonymous, as denoted by RT.

[11] UK Data Archive, "Find Data: Our HASSET Thesaurus," http://www.data-archive.ac.uk/find/hasset-thesaurus.

[12] UK Data Archive, "SKOS-HASSET Help for Users," http://www.data-archive.ac.uk/media/393116/skos-hasset-help.pdf, 3.

[13] OCLC, "VIAF: Virtual International Authority File," http://www.oclc.org/viaf.en.html.

[14] This list is comprised, in part, of the following sources: UNESCO, "Other Thesauri, Glossaries and Terminology Databases," http://databases.unesco.org/thesaurus/other.html.