eHRAF Databases: Development Plans to Improve Searching and Analysis

Carol R. Ember, Human Relations Area Files at Yale University
Michael D. Fischer, University of Kent and Human Relations Area Files

September 10, 2017

Abstract

The HRAF databases, eHRAF World Cultures and eHRAF Archaeology, each contain large corpora of curated qualitative ethnographic texts, subject-indexed at the paragraph-level by anthropologists describing social and cultural life in the past and present. In both paper and online formats, the detailed subject-indexing greatly facilitates the retrieval of information, but in most searches there are serious problems of scale, while researchers have had to largely depend on their own knowledge and ingenuity for analysis. HRAF is engaging with development plans to improve both search and analytic capabilities. Search will be improved through search by example, semantic post-processing and topic maps and models better leveraging the subject-index codes. Analytic support, such as visualization, summarization and support for researcher-directed qualitative and quantitative analysis of search results, will be accessible through web-based facilities provided by HRAF, and through an extensible services platform directly accessible by researchers using R and other research platforms. Further down the road in development terms, we plan to develop computerized auto-coding or interactive computer-assisted coding that might assist in developing post-hoc sub-categories (variables) with normalized values for analysis. Such auto-coding could also increase the potential for interoperability across databases.

Introduction

There is a great need to radically expand international capacity for secondary comparative cross-cultural research, making accessible improved exploratory and analytic methods and tools and improving the data infrastructure, which will greatly enhance and extend means for leveraging the award-winning ethnographic and archaeological online databases deployed and maintained by the Human Relations Areas Files (HRAF) at Yale University. To engage a broader constituency of researchers we are developing computational approaches that will address a broad range of theoretical and practical approaches for exploring and analysing ethnographic data in a comparative manner. This will not only increase capacity in comparative research as it has been done in the past but will greatly extend the relevance of comparative cross-cultural research to new areas and problems for which comparative ethnographic research could be applied. Most of these stem from the key driver underlying the original project by the Institute of Human Relations at Yale University in the 1930s—to bring together the knowledge we have of the world's societies for a better understanding of human behavior in all its variety.

The core data for the HRAF databases are descriptive texts—originally ethnographic and more recently archaeological. Ethnography is central to research in cultural and social anthropology as a comparative discipline dedicated to studying the diversity of humanity and its productions. The collection of primary ethnographic data is generally the outcome of fieldwork, which includes records of experience from participant observation, targeted observation, informal, semi-formal and formal interviews, collections of materials, surveys, focus groups and oral and written material produced in or about the course of daily life, such as speeches, dance, musical performances, political events, newspapers, magazines and other productions. Although historically anthropologists have been the principal agents of primary ethnographic research, ethnography as a source of data has become increasingly influential in many academic and applied areas, including education, aspects of economic and social policy, diplomacy, legal policy and legislative lawmaking, public administration, industrial manufacturing, institutional governance and military policy, among many others. Although primary ethnographic research is an important component in many of these applications, reuse as secondary ethnographic research has a much broader reach.

Indeed, generalizations about human behavior, the object of cross-cultural research, cannot be done without a corpus of ethnography to analyze. *eHRAF World Cultures* (HRAF n.d.a), with paper precursors in the 1930s and 1940s, has facilitated comparative research. With these and other resources, cross-cultural researchers have made substantial progress testing theories about a wide variety of topics using existing databases. Results of such hypotheses tests on 10 or more societies are summarized for about 800 studies on HRAF's new open-access database—*Explaining Human Culture* (HRAF n.d.c.)— we estimate that we have covered about half of the existing studies to date. (Previous summaries of cross-cultural findings can be found in Levinson and Malone 1980, Ember and Levinson 1991). In comparison, cross-archaeological research is in its infancy (for some exceptions see work by Peregrine 2017 and Peregrine, Ember and Ember: 2004, 2007); most archaeologists use ethnographic analogy to help make inferences about the past M. Ember and Ember (1995) and Peregrine (2001, 2004) discuss how to use cross-cultural data more effectively for making archaeological inferences.

The present online database, *eHRAF World Cultures*, now contains the finely subject-indexed full text of almost 6000 primary ethnographic and ethnological documents for over 310 cultures with additional detailed systematic classificatory metadata at the paragraph, page and document level added by HRAF staff analysts. At the present time, access to eHRAF World Cultures (and for the companion eHRAF Archaeology [HRAF n.d.b] with almost 100 traditions and about 2500 documents) is through two fairly conventional web applications, currently accessed by over 500 academic libraries and research centers. Both databases were voted by *Choice Magazine* (a division of American Library Association) as two of the "Top 10 Internet Resources" in 2015.

Despite the awards, we must do much more to enhance the research potential of these databases. To be sure, the *eHRAF World Cultures* database is far faster and easier to use than its microfiche and paper predecessors (*eHRAF Archaeology* has only existed online), but there is presently little one can do directly with a given set of results other than read and process hundreds to thousands of pages of text in a conventional scholarly fashion, much as one worked with the original cellulose technology.

The databases have a long history dating from the 1930s and have undergone a number of technological innovations and transformations detailed in Ember (2012) and summarized below. **This paper outlines our development plans for a new major transformation that has three major goals**: 1) to use computer technologies to expand capacity to find relevant information by researchers based on enhanced semantic and structural metadata, with and without specialized training; 2) to develop auto-classification and auto-coding services to greatly improve capacity to analyze content; and 3) to establish a services framework to support conventional and new methods of describing, summarizing, visualizing and analyzing data.

There are a range of methods for doing cross-cultural ethnographic research applied from many disciplinary perspectives, including anthropology, psychology, behavioral ecology, economics, history, or any discipline that can benefit from situating a given behavior or belief within a pan-cultural or pan-geographic context. Ethnography is not only the cornerstone of cultural and social anthropology, but increasingly important in other academic disciplines, such as political science, sociology, and education, as well as in industry and such allied areas as marketing and advertising. Our aim is to provide support for a broad range of methods, rather than to select specific methods, and thus to provide an extensible services architecture and support tools such that a researcher can script a set of operations using a set of building blocks we will provide, or in some cases, collaborate with HRAF to develop support for methods that go beyond what is available. In addition to services that search on relatively low level criteria in the database, we will provide metadata and service-level tools to improve capacity to locate relevant material at a higher level than direct keys or literal text through semantic search capabilities derived from metadata and from a ‘blunt’ model theoretic semantic representation of the text. In addition to returning literal results from the database, we are providing facilities to process these in a researcher-centric manner to support a range of researcher-modifiable classificatory models to facilitate one of the most common activities across the range of researchers, assigning codes to mark states or activities that can be used in subsequent quantitative or qualitative analyses. In addition to results from the database, in the longer term we will provide means for private ingestion, integration and search of a researcher’s material with the eHRAF databases, possibly extending these to a public archive.

Access to the enhanced eHRAF databases using associated tools will be available by membership on a cost-basis through existing and new membership plans that ensure the long-term sustainability of the data and associated tools, and meet existing copyright legal restrictions on use of ethnographic publications in the database.

HRAF Database History

Before the advent of computers, in the 1930s academics at Yale’s Institute of Human Relations, convinced that scholars should study humans in all their variety not just those closest to home, were interested in producing data about cultures of the world that could be rapidly retrieved by scholars in many different disciplines. The basic data was primarily ethnographic in nature; that is largely text information about cultural and social life based on participant observation and interviewing. The pre-computer organized information systems developed at the Institute were a technological break-though and provided a backdrop for the computerized (and now online) versions that supplanted it. In 1949, HRAF became a financially independent nonprofit membership organization that remained affiliated with Yale. Although the original “HRAF Files” were produced in paper (later microfiche, CD-ROM, then web) these contained innovations that for its time gave scholars unparalleled access to a large quantity of comparable textual and graphic information about the cultures of the world. The databases in all formats were based on the following principles (Ember 2012): 1) use original text so that researchers could make their own evaluations; 2) develop a systematic subject-classification system (*Outline of Cultural Materials* [OCM]—now with over 700 subject categories); 3) use human intelligence to subject-classify at the paragraph level and sometimes sentence-level; 4) make the materials available in one place as a discrete collection; and 5) with appropriate metadata, physically put materials on the same subject by all authors together for each culture. These innovations greatly facilitated worldwide hypothesis-testing cross-cultural research in subsequent decades (Ember 2012). Since 2014 we have been working on the next generation of our web application based on a services framework. We have made sufficient progress to plan and undertake development of new ethnographic data resources and new tools for analysing ethnographic data made possible through leveraging ubiquitous access to HRAF’s database, the largest database of curated and augmented ethnography in the world.

One of the first orders of business was to develop a topic classification system that could help scholars find similar types of information despite vast differences in custom and terminology used in different regions and cultures, including normalizing the many alternative names for most cultures. To take a simple example, all societies we know of have some kind of dwelling where families live, but ethnographers could use native terms (such as the Navajo term “hogan”) or they could use alternative words like “hut,” “house,” “tent,” “pit-house,” etc. The HRAF staff decided to create a number of different subject categories pertaining to residences. One, “Dwelling,” a subcategory of “Structures” describes residential structures with an emphasis on their physical attributes, such as mode of construction, shape and size, the durability or portability of the structures, or their seasonal uses. In contrast, the subject “Household” focuses on the social aspects of family units, such as typical and varying composition of households and whether household members live in one dwelling or a group of buildings within a compound. Other categories cover how buildings are constructed or what the interiors are like. Of interest is that the creators of the *Outline of Cultural Materials* report that they found it difficult to develop a system based on theoretical or preconceived categories; rather they noted that it was necessary to develop the system more inductively through trial and error, that is, after reading a variety of ethnographic materials seeing how anthropologists and other observers organized their materials (Murdock et al. 1950, xix). The result, the *Outline of Cultural Materials* (OCM), first published in 1938, was revised in print 12 times (6 editions and 6 editions with modifications—the latest print edition is Murdock et al. 2008). The OCM provides over 700 categories of controlled vocabulary to characterize subjects. As a shorthand, all subjects were given three-digit numbers, with the first two digits representing the more general category. (For example, Dwellings is 342, under the broader Structures category of 34*.) The OCM categories are not just used by HRAF; museums use it to classify their materials and individual ethnographers have used the subjects to classify their own field notes. Although controlled vocabularies are not unique, what is unique to HRAF is the fine level of subject-indexing to the search and retrieval elements (SREs—typically paragraphs). The other classification system, the *Outline of World Cultures* (OWC) provides a standardized list of the cultures of the world to resolve the problem that the same group may be referenced through many different names; cultures were given alphanumeric identification numbers, generally reflecting the region and country location of the culture. (The first print edition appeared in 1954 [Murdock 1954] and the 6th edition in 1983.) The HRAF staff concluded that some of Murdock's regions were problematic, particularly the grouping of Muslim cultures together as "Middle East," even though many were in sub-Saharan Africa. Therefore, in *eHRAF World Cultures* new broader terms for region and subregions were introduced that were based more on geography.

The present database consists of the full text of 8000 ethnographic and archaeological documents, represented as approximately 3 million paragraphs together with photographs, tables and other documentation. The content of each paragraph is augmented with topical codes assigned by our team of analysts against the specifications in the *Outline of Cultural Materials*, a thesaurus that HRAF continues to develop. The societies and cultures covered are normalized using the *Outline of World Cultures*, as these are given many different names in the original sources. The development strategy up to now has been to include societies documented in a good range of different sources over a period of time, so that both synchronic and diachronic analysis can be supported. Since the societies are selected on a criteria of data quality and quantity, and because societies are added and augmented annually, the collection as a whole does itself not constitute a statistical sample. However, the current collection currently includes subsamples (the Probability Sample Files and a Simple Random Sample) that can support statistical tests. And, in a few years we will have added all the Standard Cross-Cultural Sample societies (as of 2017 contains about 80% of that sample). While the present web based application currently used to access the database is many orders of magnitude faster and easier to use than previous versions, the methods of use and analysis it supports are pretty much identical to those of the early paper-based versions of the database. The main advantages relate to time consumed in retrieval and handling of results, but there is little that can be done with the results beyond storing and reading these, assessing, classifying, or coding for specific issues. Of course, the great reduction of time and difficulty in handling of results does mean more research and more ambitious research can be undertaken. However, in practice the original methods do not scale upwards well enough to realise the full benefits.

On this basis membership has grown from an initial 5 member institutions to over 500, sufficient to support modest database development and maintenance on a self-funding basis. Over the past three years we have envisaged and specified a development path to expand research capacity using the database. We do not, however, have sufficient resources to develop the next generation of tools and enhance the database to support a much broader range of cross-cultural research at a time when many of the problems of globalization could be better addressed by more information about the past and present of the different parties involved.

The HRAF files greatly facilitate qualitative and quantitative comparative research, especially compared to the time it would take to collect all the books, articles and manuscripts and then find relevant material, but the scale of the of data returned in typical search results is often problematic because we do not have methods that scale accordingly.

Although the size of the HRAF collection is not extraordinary with respect to some 'big data' datasets, just a few gigabytes, the structure is heterogeneous and complex and presently most of the relevant information to be used in research must be extracted from ordinary document text through reading and assessment for the intended research objectives.

The capacity to collect together relevant text fragments from multiple sources has greatly expanded the capability of researchers to do meaningful comparative cross-cultural research. However, a range of better search facilities and post-processing tools and methods for the returned text will expand researcher capacity, and thus make detailed comparative and cross-cultural research attractive to a wider range of researchers within and outside anthropology.

Expanding HRAF Research Services

HRAF's IT team has already taken steps that move towards the outlined goals. The first step, currently underway, is to restructure the underlying XML data. Essentially, we are repurposing search and retrieval operations as a variety of services that are independent of any specific web application. In contrast, our present HRAF XML schema is oriented towards reproducing the original appearance of a publication, and has a very complex structure due to the heterogeneity of the 8000 or so sources we use, spanning over 100 years of ethnographic evolution. We plan to retain this structure for production and archival purposes, but to facilitate large scale search and retrieval services we are normalizing the structure to focus more on associating key metadata and pre-compiled statistics with each paragraph so that it can be more easily evaluated within a given search, and a broader range of search criteria used. In addition, we are working on a new application based on a services architecture framework, so that new features can more readily be added.

The new services will leverage attributes that identify pertinent metadata such as culture, region, time of description, time of publication, type of author (e.g., ethnologist, geographer, missionary), in addition to the present text and associated analyst supplied OCM subjects. We will develop auto-classification capabilities based on the HRAF collection that will enhance conventional topic extraction (ontological classification), tools to support coding materials for value (epistemological assignment) and advance auto-coding techniques to promote broad consideration of comparative analysis and situation of human practices and behaviors for basic and applied research. To support researcher initiated data-mining we will develop services with which we will experiment with different approaches to producing topic models and topic maps suitable for paragraphs in context and with auto-classifying paragraphs and larger sections with OCM categories in a manner consistent with our professional analysts. We will also explore methods to apply similar auto-classification to non-ethnographic sources, ranging from academic publications in anthropology to newspaper articles. In the future, we will work towards expanding access to and reuse of other researchers' contributed underlying and published ethnographic and other data, without compromising confidentiality or other constraints, to promote reuse of data generally in a new services platform that will enable many researchers to add their own materials to enhance the ethnographic corpus and promote re-use of ethnographic data within the bounds of well-established and well-founded ethical constraints.

Reuse of Ethnography

In a May 2009 four-field workshop we framed the argument for developing a strategic plan for the digital preservation and access (DPA) of anthropological research materials. (See Ember et al. 2009 for a report.)

Our proposed research and the computer-assisted tools we develop will not only support conventional reuse, and eventually DPA (through researcher deposits), but new types of research. In the course of our work, we will be examining very large corpuses with topic "extraction" techniques and address the following research questions: How have topics in ethnography changed over time? How do topics vary by region of the world? How do different types of fieldworkers vary? Do topic extraction techniques produce similar results on the field notes and the formal ethnography produced by the same ethnographer?

These analyses will be the basis of several publications. In addition, the computer-assisted tools we develop will aid researchers in their current research. For example, the tools will help them look for similar topics throughout a large set of notes (their own or others'), collect disparate information contributed by the same informant, and creating project-specific indexing systems. A further goal is to develop robust topic identification that could facilitate comparisons across corpuses, multiple repositories and archives.

The proposed development uses a combination of mostly relatively stable technologies and some technologies that are novel. However, the proposal work packages are designed so that the objective of each work package can minimally be achieved though tested technologies. For some work packages there are additional elements that we have high confidence in, but can only be established as valuable once implemented. These more risky elements are in the work packages relating to individual application services rather than the core. The services framework proposed is a novel approach, but we have already done enough work with it to argue it is simply novel, but not risky. It arose directly to serve the precepts and requirements of comparative cross-cultural research using the present HRAF database.

Cross-cultural researchers have made substantial progress testing theories about a wide variety of topics (see *Explaining Human Culture.).* Almost all these studies are synchronic and yet we know that theories that postulate causal sequences should be tested with diachronic data, which is possible for most of the cultures in the eHRAF databases. The new metadata for the proposed development should facilitate that goal inasmuch as "time of description" will be a more prominent feature of result sets. This temporal metadata should also facilitate investigations of historical changes in the ethnographic corpus, e.g. how ethnographic descriptions have changed over time. In addition, the archaeological database clearly can be used for testing diachronic hypotheses. A recent study by Peregrine (2017) demonstrates how *eHRAF Archaeology* can be used to test causal hypotheses related to the impact of climate-related disasters on cultural attributes.

These new services will greatly facilitate measuring and coding more abstract and conceptually more difficult cultural constructs. For example, in a current grant project (that Ember is leading)we are coding "tightness" and "looseness" of cultures to test the theory that natural hazards increase cultural tightness (referring to the pervasiveness of strong norms and punishment for deviance—Gelfand et al. 2011). Coders are currently reading material across five domains of culture (marriage, sexuality, gender roles, socialization, and funerals) as well as reading about norms and punishment. While the current OCMs narrow down material somewhat, the relevant passages from eHRAF often exceed 300 pages, making the coding task very onerous. Summarization, visualization, topic maps, agent/action identification and textual metrics services we develop will help in arriving at much more efficient ways to view and code the data, greatly expanding the opportunities for research.

Goals and Overview

To achieve our three major goals (see 6th paragraph of the introduction) we will ***first*** expand the resolution of metadata in two important ways: use HRAF's subject classification at the paragraph-level--using the *Outline of Cultural Materials* (OCM) in conjunction with text-processing, topic extraction, natural language processing (NLP) and other methods based on analysis of the HRAF corpus to enable finer contextual distinctions; and, make better use of existing document and page level metadata, such as date of fieldwork, time that the cultural information pertains to, and location references. The metadata will also include the type of author (e.g., ethnographer, geographer, missionary). The expanded metadata will improve relevance of result sets and make possible much finer grained extraction of information from a given subset of textual results in a form suitable for further qualitative or quantitative analysis using appropriate services. The time and place metadata will facilitate diachronic comparison and evaluation of intra-cultural variation. ***Second,*** we will develop auto-classification services that assign subject metadata consonant with categories from HRAF's *Outline of Cultural Materials* at the paragraph-level to enable use of private material a researcher might ingest for their own use. We will also explore methods to apply similar auto-classification to non-ethnographic sources, ranging from academic publications in anthropology to newspaper articles. ***Third,*** to improve research capacity of the database we will restructure the database into an extensible framework based on virtual documents (HRAF vDoc Services Framework —see appendix 1) coupled with an initial set of new analytic services to expand processing of results. In connection with this we develop a services framework with a flexible architecture for mounting services to support a range of new methods for summarization, description, analysis and visualization of query results, ***Fourth***, we will deploy a workflow manager that facilitates management and further processing of result sets, including subsequent transformations and analysis history; this “manager” will facilitate the ability of researchers to create and document complex workflows with relatively little need to understand the services framework itself. In the future we will work towards expanding access to and reuse of other researchers’ contributed underlying and published ethnographic and other data, without compromising confidentiality or other constraints, to promote reuse of data generally in a new services platform that will enable many researchers to add their own materials to enhance the ethnographic corpus and promote re-use of ethnographic data within the bounds of well-established and well-founded ethical constraints.

Development Plans

To achieve these goals we will:

a. Substantially expand and revise the metadata for HRAF’s extant databases for the purpose of supporting new methods and forms of research leveraging ethnographic data, either single culture or cross-cultural research involving representative samples of a regional or global nature. i) This will include transformation of the OCM into a RDF semantic graph rather than its present form as a thesaurus. ii) Leveraging this view of the OCM in context with the OCM augmented text in the database to produce topic models and from these focused topic maps based on code libraries compliant with the ISO/IEC JTC1/SC34 Topic Maps Data Model standard and XTM2. iii) Applying automated linguistic analysis based on libraries relating to the Stanford Parser (SP) and WordNet to create metadata identifying part of speech, parsing using SP for subject, object, predicate, modality, tense and negation. These will be used to develop a ‘blunt’ model theoretic semantic model (Thomason, 1974) of the of the entire text, corresponding to a formal ethnographic framework proposed by Bock (1986) that describes relations between different kinds of agents, and describes potential social interactions and relations. This means that a segment of retrieved text can be ‘executed’ to create semantic ‘side effects’ representing agency-based processes and states of affairs from the relationship of the different logical phrases in the text (Fischer 2006; Horty 2001) and possibly the original textual context of a given text fragment. A deontic interpretation (e.g. Castro and Maibaum 2007; Dong and Li 2013) of these side effects, which focuses on the range of potential outcomes and their relationship rather than predicting what will happen, effectively can help characterize activities or states of interest to researchers (arranging a marriage, engaging in a ritual, performing violence) that can be used to select text candidates for further interrogation, support auto-coding, support the construction of topic maps and support writing summaries of larger text segments. This will be generalized semantic account to some extent, but should be within the scope of what ethnographers are usually trying to identify when they read a text. This will support search and extraction of specific data from text, summarization, and production of meaningful textual metrics, as well as more complex constructions as identifying Bock’s actor-action-actant relationships and contextual environments for interpretation in the text. These relationships will be used to construct RDF graphs in conjunction with the topic map that are manipulated through application services. Also supports coding, manual and auto, in general.

b. Based on the expanded metadata in a) to develop means of auto-classification for a range of genres of ethnographic texts and relevant descriptive material that researchers may need to incorporate for their research. This will take two forms, one auto-coding and a more specialised auto-coding relating to predicting OCM codes for a text fragment. The former supports an activity at the core of most current hypothesis driven comparative research, the identification of values for specific variables of interest to the researcher. This is usually achieved through reading the text. We will start with algorithms such as TF–IDF, which has used in other projects, such as identifying illegal wildlife sales online and finding similar articles. We will then experiment with other algorithms. This case is rather more complex, as interesting variables often relate to activities and outcomes rather than simple structural criteria. Leveraging the model theoretic representation in a) which generalizes types of agents and activities, together with other metadata we will create a mode of search and identification based on a model representation of what kinds of processes and agent relationships the researcher is interested in, and to match these to the accounts that can be derived from the text. While this is only practical when processing a set of results that has been retrieved using more conventional search criteria, it opens up possibilities for reducing the result set to more pertinent results from what might be a few hundred to a few thousand pages of text for a given query.

c. Develop analytic tools for assigning and using metadata with researchers' private external data sources they wish to compare/contrast with content in the eHRAF collections. Although many researchers will be content to rely on secondary research, many will want to evaluate specific material they have with respect to the larger population in eHRAF. Using auto-classification and the facilities and services used to fulfil a) and b) above, we can apply this to private material which can be kept private, but benefit from inclusion from the researcher's point of view as a part of the dataset they are analyzing. The ultimate goal is to assign OCM codes to arbitrary ethnographic, and other, texts which could then be ingested into an auxiliary dataset that can optionally be included when searching the HRAF database.

d. Developing analytic tools for 'blind' sources, which are either proprietary or very sensitive in nature. This means that these sources can yield a great deal of information, without revealing the details underlying text. This will permit us to begin a 'searchable' archive of ethnographic field notes, which are generally not accessible to other researchers. In addition to textual metrics of the text, applying the methods in a) and b), this will include summaries, visualizations, non-specific paraphrasing (where names, and in some cases places, are omitted), 'blunt' model-theoretic semantics using Bock's framework and identifying similar examples drawn from the HRAF database.

e. Develop analytic tools for assigning and using metadata with external data sources, either public web servers or with data collaborators. The concept is that we would source a Virtual Document (vDoc—see below) from an external source, where it could then be manipulated and analyzed using the remainder of the services framework. We are not expecting it to be quite that easy, particularly for web sources that are not documents per se, and certainly not ethnographic ones. However, we will have learned a good deal about different genres of document by the time we get to this work package which should help implementation. We are confident that we can get this working, but uncertain as to the usefulness of the results on unsuitable material.

We are restructuring the HRAF databases to better support research. The present database's XML components are converted to the vDoc schema. This is done in eXistDB using an XQuery script, and a preliminary specification for vDoc has been implemented although we are still making modifications. This specification will continue to require minor adjustments and additions, but vDoc is designed to be extensible so it can be modified to reflect new metadata and content types. However, any such extensions will be backward compatible. For practical reasons the base data will remain in our production XML schema, and the vDoc instance refreshed from that instance as needed, so that we remain synchronized to the growing HRAF corpus. We only need to refresh new documents, and, rarely, modified documents.

We are thus establishing an extensible HRAF vDoc Services Framework (HRAFvSF). A vDoc is a container used to store and reference a result set of text. Any vDoc can be dynamically instantiated and searched and further processed, with any results populating a new vDoc. The services framework itself is fairly conventional based on node.js, which can also pass on services based on other code. A task manager will be created so that a simple user interface can be produced for researchers to originate, manage and further process results, tracking the history of operations on the vDoc contents.

Based on this framework we will establish core services, component services for vDoc pipelines (see Appendix). The simplest pipeline would have a reference vDoc as embedded source, and a renderer (serializer) that results in a user view based on the vDoc. More complex pipelines will use source references and process the source (or intermediate transformations) in various ways using core and application services. Pipeline services will consist of one or more of sources, transformers, aggregators, serializers, and logical operations such as conditionals, and a lambda operator for existing pipelines to substitute one or more parameters. The task manager will help end users manage these to create and track interactive and focused research.

Application services will include data extraction, pattern identification, visualization and summarization. Application services are services that perform complex operations on one or more set of data references wrapped in vDoc structures. An application service could simply be a predefined vDoc pipeline, or may use a number of programming languages to perform operations on the data in a vDoc. Application services will perform relatively high level operations: summarization, a pattern visualization, follow semantic paths through a graph and aggregating the results, finding similar texts, filtering texts for particular combinations of topics and OCM codes, producing a report for the user, storing a vDoc in a persistent store and many other operations. There is no end to this work package per se, but we expect to have many useful services in a few years.. Many of these will be in the form of pipeline scripts which can serve as reference examples for researchers.

We will enhance HRAF databases with extracted topic metadata use topic extraction techniques (“text mining”) to create finer-textured metadata than the OCM codes, while still leveraging the OCM codes. Text mining refers to the use of computational methods to extract significant terms and relations between terms from segments of text. Textual words are compared with those from a larger corpus of texts to derive measures of similarity. (The most common method is to use vector comparison methods, such as cosine similarity. These are tuned by transformations, stemming, and other techniques to find similar text segments by the closeness of the match.) Based on the expanded metadata produced, we will develop a service for auto classification of a range of genres of texts that authors, researchers or publishers contribute. This expands the reuse of data that would otherwise be difficult to incorporate into a study. We will leverage the OCM in context with the OCM augmented text in the database together with topic models extracted using algorithms such as LDA and LSA to produce focused topic maps based on code libraries compliant with the ISO/IEC JTC1/SC34 Topic Maps Data Model standard and XTM2. These relationships will constitute a separate dataset that points into the base vDoc instance.

Appendix 1: HRAF vDoc Services Framework (HRAFvSF)

We plan a flexible front end for service deployment for services conforming to one of many possible APIs. Inspired by the “Software Tools” concept defined by Brian W. Kernighan and P.J. Plauger in 1976. Sources and services can be pipelined, with the framework handling conversion between APIs. Internally, all requests are formatted as vDocs, a HRAF designed container for sets of content items. The sets may be comprised of literal text items, or more often the contents are vRefs—references to text items - or the set contents are defined by one or more source vDocs, a query and/or combination of queries, a pipeline aggregator that combines sets into a single set, a pipeline transformer that modifies sets, a pipeline rendering service that conditions results, a reference to another vDoc or vDoc subset, a reference to another vDoc pipeline, a reference to a binary object, or service defined object, together with metadata, a core of which is standard across vDocs and internal vRefs, with additional metadata depending on the source of the vRef or vDoc. All queries are ultimately against a collection vDoc that contains all the literal contents of the database, but all queries can be performed against any vDoc, regardless of whether the contents are references to the collection vDoc, other vDocs, queries, RDF triples, literal text or other contents. Results can be returned either as a vDoc reference, or an instantiation of the returned vDoc reference, dependent on the original request. Results that are vDocs will normally contain a history of where and how the contents of that vDoc result set were derived.

Although the HRAF-SF is designed to be extensible to a wide range of service request types to leverage mixed ability in research groups using HRAF, to leverage existing service instances, to encourage third party contributions of service definitions, the proposed work will use a more limited range of service APIs, mainly the vDoc API defined by HRAF transported using the REST API. Coding for the framework and services used will be mainly in C++, Java, Python and Javascript.

HRAF has created an initial specification of HRAF-SF and vDocs, although these will be extended. HRAF has developed some services (bibliographic, a user notebook) to production and are currently in use or will be by early 2018. More services have been developed as prototypes, including a search service, a topic map service, summarisation services, statistical services, bag-of word services etc. These will need considerable development for production use, but the algorithms are basically sound and in use to support two 'big data' research projects at other institutions by members at Harvard and Oxford. Non-service utilities have also been developed that can transform or collect data across an arbitrary range of the collection. Other code, particularly relating to pipelining operators, has been developed for another project (Cook Islands Biodiversity and Ethnobiologial Database) and will be adapted to this effort.

References

Bock, Phillip K. 1986. *The Formal Content of Ethnography*. International Museum of Cultures.

Castro, P. and T. Maibaum. 2007. "A complete and compact propositional deontic logic." *Theoretical Aspects of Computing–ICTAC*. Springer Berlin/Heidelberg. 109-123

Dong, Huimin and Xiaowu Li 2013. A Deontic Action Logic for Complex Actions. Logic, Rationality, and Interaction. Lecture Notes in Computer Science Volume 8196, pp 311-315

Ember, Carol R. 2012. Human Relations Area Files. In *Leadership in Science and Technology: A Reference Handbook*, vol. 2. William Sims Bainbridge, ed. Los Angeles: Sage Reference, pp. 619-627.

Ember, Carol R., Eric Delson, Jeff Good, and Dean Snow. 2009. Toward an Integrated Plan for Digital Preservation and Access to Primary Anthropological Data. AnthroDataDPA: A Four-Field Workshop. Chair Report. http://anthrodatadpa.org

Ember, Carol R., and David Levinson. 1991. The substantive contributions of worldwide cross-cultural studies using secondary data. *Cross-Cultural Research* 25, no. 1-4: 79-140.

Ember, Melvin, and Carol R. Ember. 1995. Worldwide cross-cultural studies and their relevance for archaeology. *Journal of Archaeological Research* 3, no. 1: 87-111.

Fischer, Michael. 2006. Cultural Agents:A Community of Minds. Engineering Societies in the Agents World VI. Dikenelli, O., Gleizes, M. and Ricci, A. (eds). *Lecture Notes in Computer Science* 3963, Springer- Verlag (2006), 259-274.

Gelfand, Michele J., Jana L. Raver, Lisa Nishii, Lisa M. Leslie, Janetta Lun, Beng Chong Lim, Lili Duan et al. 2011. Differences between tight and loose cultures: A 33-nation study. *Science* 332, no. 6033: 1100-1104.

HRAF. n.d.a. *eHRAF World Cultures*. Human Relations Area Files. http://ehrafworldcultures.yale.edu

HRAF. n.d.b. *eHRAF Archaeology*. *Human Relations Area Files*. http://ehrafarchaeology.yale.edu

HRAF. n.d.c. *Explaining Human Culture*. http://hraf.yale.edu/ehc

Horty, J. 2001 *Agency and Deontic Logic*. Oxford University Press, Oxford.

Huimin Dong and Xiaowu Li 2013. A Deontic Action Logic for Complex Actions. Logic, Rationality, and Interaction. *Lecture Notes in Computer Science* Volume 8196, pp 311-315

Kernighan, Brian W. and P.J. Plauger. 1976. *Software Tools*. Boston: Addison-Wesley Professional.

Levinson, David, and Martin J. Malone. 1980. *Toward explaining human culture: A critical review of the findings of worldwide cross-cultural research*. New Haven, Conn.:HRAF Press.

Murdock, George Peter. 1954. *Outline of World Cultures*. New Haven, Conn.: Human Relations Area Files.

Murdock, George P. 1962ff. Ethnographic Atlas. *Ethnology*. [Note: in various issues from 1962 onwards.]

Murdock, George Peter. 1983. 6th revised edition. *Outline of World Cultures*. New Haven, Conn.: Human Relations Area Files.

Murdock, George P., Clellan S. Ford, and Alfred E. Hudson. 1938. *Outline of Cultural Materials*. New Haven, Conn.: Institute of Human Relations, Yale University.

Murdock, George P., Clellan S. Ford, Alfred E. Hudson, Raymond Kennedy, Leo W. Simmons, and John W.M. Whiting. 1950. *Outline of Cultural Materials*. New Haven, Conn.: Human Relations Area Files.

Murdock, George P., Clellan S. Ford, Alfred E. Hudson, Raymond Kennedy, Leo W. Simmons, and John W.M. Whiting. 2008 with modifications. *Outline of Cultural Materials*. New Haven, Conn.: Human Relations Area Files.

Peregrine, Peter N. 2001. Cross-cultural comparative approaches in archaeology. *Annual Review of Anthropology* 30, no. 1: 1-18.

Peregrine, Peter N. 2004. Cross-cultural approaches in archaeology: comparative ethnology, comparative archaeology, and archaeoethnology. *Journal of Archaeological Research* 12, no. 3: 281-309.

Peregrine, Peter N. 2017. Political participation and long-term resilience in pre-Columbian societies. *Disaster Prevention and Management: An International Journal* 26 (3: 314-329

Peregrine, Peter N., Carol R. Ember and Melvin Ember. 2004. Universal patterns in cultural evolution: An empirical analysis using Guttman scaling. *American Anthropologist* 106, no. 1: 145-149.

Peregrine, Peter N., Carol R. Ember, and Melvin Ember. 2007. Modeling state origins using cross-cultural data. *Cross-Cultural Research* 41, no. 1: 75-86.

Thomason, Richmond H. (ed.) 1974. *Formal Philosophy*. New Haven: Yale University Press.