Big Data and Research Opportunities Using HRAF Databases

Michael D. Fischer
Affiliation: University of Kent, Canterbury, UK and the Human Relations Area Files at Yale University
Email: m.d.fischer@kent.ac.uk
Phone: +44-771-357-4737

Carol R. Ember
Affiliation: Human Relations Area Files at Yale University
Email: carol.ember@yale.edu
Phone: +1-203-764-9401

Abstract

The HRAF databases, eHRAF World Cultures and eHRAF Archaeology, each containing large corpora of curated text subject-indexed at the paragraph-level by anthropologists, were designed to facilitate rapid retrieval of information. The texts describe social and cultural life in past and present societies around the world. As of the spring of 2017, eHRAF contains almost 3 million indexed "paragraph" units from over 8,000 documents describing over 400 societies and archaeological traditions. This chapter first discusses concrete problems of scale resulting from large numbers of complex elements retrieved by any given search. Second, we discuss potential and partial solutions that resolve these problems to advance research, whether based on specific hypotheses, classification or identifying and evaluating embedded patterns of relationships. Third, we discuss new kinds of research possibilities that can be further advanced, have not yet been successfully attempted, or have not even been considered using anthropological data because of scale and complexity of achieving a result.

Introduction

The database known as the "HRAF Files" in paper was for its time (in the 1930s and 1940s) a technological breakthrough giving scholars unparalleled access to a large quantity of textual and graphic information about the cultures of the world. This paper briefly discusses the initial innovations, the enhancements with online searching (now eHRAF World Cultures and the newer eHRAF Archaeology), and future developments planned. As of the spring of 2017, the two eHRAF databases contain almost 3 million "paragraph" units from over 8,000 documents describing over 400 societies and archaeological traditions.

Before the advent of computers, 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. It involved the following principles (Ember 2012): 1) use original text so that researchers could make their own evaluations; 2) organize the materials by systematically classifying subjects as well as cultures; 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) physically put materials on the same subject by all authors together for each culture.

The first order 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 world; 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 actual process of producing the "files", originally called the Cross-Cultural Survey (later referred to colloquially as the "HRAF files," but we use the more appropriate name, HRAF Collection of Ethnography), was extremely labor intensive because it predated not only computers but also copy machines. So, each paragraph had to be retyped using onion-skin paper and carbon paper. If a paragraph was about more than one subject it had to be duplicated as many times as it had subjects because of the need to put all the same subject together. Because of duplication needs, the paper version of the files had about 4,000,000 pages of information.

The dilemma in any searching method, whether using an analog or digital means, is how to balance the breadth of coverage (i.e., the number of retrieved elements) with the efficiency of a smaller search result. Even with pinpointed OCM subjects, the number of results can be quite large. For example, the category "Dwellings" yields almost 20,000 paragraphs for 292 cultures in eHRAF World Cultures; "Households" has almost 40,000 paragraphs. A few strategies can help narrow the scale. For example, cross-cultural researchers generally test specific hypotheses on smaller samples of societies claimed to be representative and limit their reading for each society to a focal community in a particular time period (Ember and Ember 2009: 76-78). Narrowing to a focal community and time period usually means that only those documents pertaining to the right foci are perused. HRAF has provided aids for this, such as marking the documents that match the foci for the Standard Cross-Cultural Sample, a commonly used cross-cultural sample (see http://hraf.yale.edu/resources/reference/sccs-cases-in-ehraf/). Specific hypotheses usually mean that a researcher can narrow the scope of a search to fewer paragraphs. So, for example, a researcher may be interested in the size of dwellings and how size varies with different aspects of social structure (Divale 1977, M. Ember 1973, Porčić 2010, 2012).

Another strategy (possible in the online eHRAF databases) is to narrow searches by combining subject categories or narrowing by adding keywords. For example, if you want to know about family household and dwellings and you believe they are likely to be described in the same paragraph you can ask for both categories in the same advanced search (this narrows the number of paragraphs to about 1500 paragraphs). Adding keywords to the search works well if there are only a few commonly used distinct words or phrases. If you are interested in the size of dwellings you can add the keywords "feet" or "meters" to narrow the Dwelling subject search. However, keyword searching is problematic when too many multiple terms describe the same construct, making it almost impossible to include all the appropriate words.

Also problematic is when a subject of interest does not fit neatly into one or two OCM subject categories. In a current project, we (Ember and colleagues) are trying to measure the "tightness" or "looseness" of cultures from ethnographic descriptions. This is a broad concept that involves assessing the degree to which there are strong and pervasive norms as well as expected punishment for norm violations (Gelfand et al. 2011). Many broad subject domains have to be examined (e.g., offences and sanctions, norms, sexuality, marriage, gender, socialization) and the volume of material is so large that researchers sometimes have to spend

a week reading the material for each society before they can assign specific values to the various “tightness/looseness” measures (a process called 'coding' in cross-cultural analysis).

Although 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, the quantity of data returned in search results is still often problematic. It is relatively easy to retrieve relevant text, and being able to collect together all relevant text from multiple sources greatly expanded the capability of researchers to do meaningful comparative cross-cultural research. However, HRAF is embarking on using 'big data' methods to develop a range of post-processing tools and methods for the returned text to expand researcher capacity once again, and thus make detailed cross-cultural research attractive to a wider range of researchers within and outside anthropology. 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 most of the relevant information must be extracted from ordinary document text.

Addressing Problems of Scale in New Ways

As indicated above, one of the main problems that scholars face is having too much material even when it is narrowed by OCMs and/or keywords. HRAF is currently developing a system where researchers can store a personalized set of preliminary results in one or more “notebooks,” that can be returned to as often as needed. An initial search for a subject may be large, but the notebooks will have additional tools of selection besides deleting or adding paragraphs or adding keywords to refine a search. These include: 1) searching within collected materials in personalized notebooks with topic maps and summarization services; 2) after identifying some critical paragraphs using computer algorithms to find “paragraphs like this”…; 3) developing computerized auto-coding or interactive computer-assisted coding that might assist in developing post-hoc sub-categories (variables) with normalized values for analysis .

One approach to improving this situation that HRAF is experimenting with is leveraging the OCM classification to produce more nuanced topic-maps for the documents that can be used not only to expand capacity to search for information, but also to interpret the information within. The OCM was developed to support particular approaches to research,

and has, indeed changed over the years to reflect changing research priorities. However, it is more pragmatic than theoretical in design, beyond the broad theoretical principle of comparison that stimulated the original collection of the data. The categories used in the classification represent the broad topics that anthropologists and others have found productive, based on the theoretical and practical aspects of the ethnographic literature and its applications. Topic-mapping is local to whatever specific collection of ethnography it is applied to, and will vary depending on the corpus. A topic map of the entire collection will be different from topic maps of individual documents (or groups of documents), which will be different from topic maps derived from the results of a search. These differences can be leveraged to identify the gravity of particular topics at the different levels of mapping (terms gain 'gravity' when these also appear in prior and/or subsequent search units).

Among other things, topic mapping improves results from searching for keywords, but also helps identify sections that are strongly correlated with a keyword. For example, in Figure 1 a query for ‘oxen’ is made in Paul Stirling’s field notes (not currently in eHRAF, but smaller in scale and useful for developing services for dynamically identifying topics) in an application which returns results based on topic maps associated with the English term ‘oxen’. This expands the results from an initial 5 instances to 65 instances, including several instances where Turkish terms for oxen are used. Of the 65 results the majority (52) were directly relevant to the search term, although the term did not actually appear in the text, and the remainder were of secondary relevance in the context of the whole search results in that they answered questions that arose from the other results relating to land usage, alternatives etc. The additional 60 notes are produced because there are topical relations in the oxen notes that can be satisfied by these additional notes. As the topic map in Figure 1 shows, these contextualize the specific notes relating to oxen. Although primitive, this illustrates the potential applications of text mining for secondary research from archived resources.

Figure 1. Topic Search for Oxen in Paul Stirling's Fieldnotes

Topic-mapping, combined with the returned text and publication metadata, can also assist in creating narrative or structured summaries of a set of search results, where different parts of the summary link back to the search results responsible for that part of the summary. This leads to the prospect of a 'search by abstract', where a single abstract is produced for each search , and subsequent searches can take place based on selected portions of an abstract which retains a record of the sources contributing to each portion. Related, but more easily achieved, topic maps in conjunction with OCMs can support a 'search by example' approach, where selected entries are used as a basis for identifying similar entries from across the database.

Another development planned is the ability to do some data mining of the eHRAF corpus. We will develop topic extraction techniques ("text mining") interactively, that is, in conjunction with directed research questions. 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.). This can be further improved through leveraging the OCMs, which make it possible to 'mine' specific categories in the collection, and by using topic maps, either those local to each OCM, or collectively.

Why is interaction important? Although general ideas about the content in terms of contextualized "themes" can be "automatically" derived, these just provide an overview of themes and their relationship to other themes. Research inquiry usually requires more directed, contextualized searches. Our approach is to work from these more general topic maps to identify segments of text (paragraphs, sections, pages, user selections, etc.) that the researcher identifies as pertinent, to classify and transform these into a pattern that can be matched against other text segments for similarity. Effectively, we leverage the intelligence of the researcher to identify textual environments of interest; we can then identify similar segments without needing a high-level semantic interpretation. The researcher can also refine the search by pointing to subsets of passages that are most relevant. False positives are possible, but the aim is to reduce the matches to a manageable number. Results may be improved by the use of stemming algorithms to reduce words to their core part—for instance, "food," "foods," and "foodstuffs" become "food."

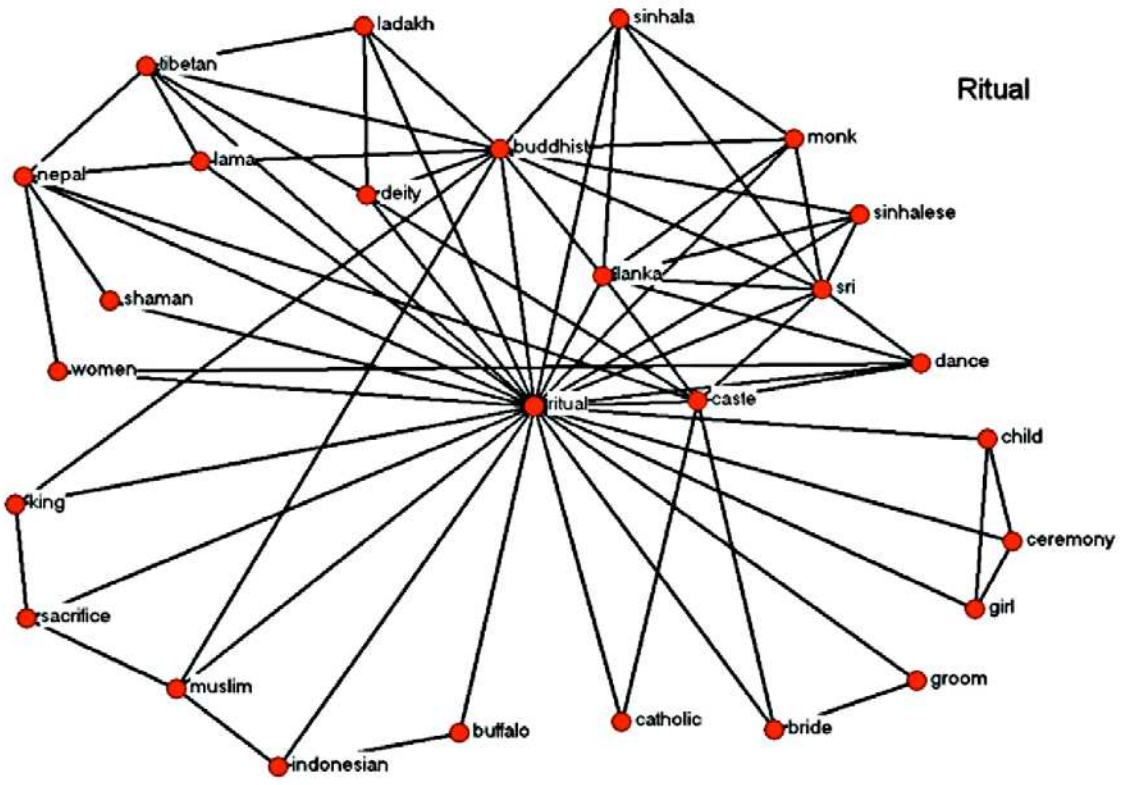

Fig. 2. Concept Network for Ritual in a sample of American Anthropologist articles 1940-1960. Nodes can be clicked to reveal source segments.

We have applied this approach successfully to text mining academic articles from the *American Anthropologist* (AA) for 1950-2000. Figure 2 shows examples of topical domains from the AA produced in a prior EPSRC/ESRC project, Genealogical Relations of

Knowledge (GROK). However, articles are highly contextualized around a topic, have fairly clear sections that can be identified, and progress thematically. As we discovered in recent exploratory work, ethnographic monographs, although formally having general sections, cover many topics repeatedly throughout the text without distinct sections or progressive development, which is even more characteristic of fieldnotes. Fortunately, we have developed an algorithm for segmenting ethnographic texts that appears to overcome some of these problems by looking for "runs" of basic relevant patterns in paragraphs within the text and then applying the previous algorithms to these newly derived segments.

With a text-mining resource we could ask questions such as: How has topical interest changed over time? Do different types of authors describe different things? Do subjects described vary by region of the world? If we added additional metadata to the text search and retrieval elements, such as the gender of authors, their nationality, their research training, then we could ask: Do male and female ethnographers describe different topics? How does nationality of ethnographer affect areas described? Does research training matter?

Moving Forward on Re-use of Older and New Ethnographic Data

Databases like eHRAF World Cultures are aimed at facilitating re-use of largely previously published ethnographic information about the cultures of the world, but HRAF, which relies primarily on institutional memberships, has only been able to produce a subset of the ethnographic information potentially available. A major limitation is the costliness of human subject-indexing to the paragraph-level. There is much more information out there--published, unpublished, and on-going—that could be studied. Frankly, the re-use value of ethnographic data is high but hard to achieve. There are efforts, which we will describe shortly, that we will try to undertake at HRAF to be able to more efficiently process materials and maximize re-use of existing material, but we also need to enable others to process their own ethnographic materials (see "Towards a Services Platform for Broader Re-use") if we are ever going to be able to scale-up.

Before describing those efforts, we will discuss some general issues regarding the goal of putting disparate ethnographic materials together. The main issue is arriving at interoperability across varied ethnographic sources and other kinds of data, quantitative and

qualitative. Ethnography, as a form of reporting physical and cultural data relating to societies, was founded around strong comparative principles and data collected across populations and/or societies. Thus, ethnography was always intended to have considerable value for re-use purposes as scholars and policy makers ask broad or narrow questions about similarities, differences, and changes in human populations. But although the goal was broadly comparative, the heterogeneity of data across time and within and across the various cultural branches of knowledge, is considerable and interoperability presents considerable challenges. Understanding the range of data and the logical possibilities for interaction is critical and yet is under-addressed in the literature.

Broadly speaking, there are three commonly understood types of interoperability: 1) structure and/or format conventions - the form in which data is represented (syntactic); 2) meaning - what does the data represent and why and how does it represent it (semantic); and 3) what we can do with a data set - the models and/or interpretations a data set supports and how these are constructed (pragmatic). After explaining these types of interoperability, we outline some steps HRAF is taking to increase interoperability.

Syntactic interoperability requires that we can identify individual records, groups of records (if any) and their relationship, and individual data items and the encoding of data items. Any metadata associated with individual data items or groups of items would be included. The interpretation of items or records is not an aspect of the syntactic specification. Specific algorithms to do many of these syntactic conversions are becoming widespread. However, much irreplaceable legacy research data is buried in unsupported, and often one-off, file formats and data layouts. Development of more general tools and services that "open" these for future use will make valuable and often irreplaceable data accessible, and help ensure future pathways to automate "rolling over" data from format to format as digital infrastructure inevitably changes. Likewise media dependencies, for example, legacy data held on media no longer well supported (e.g. paper, card decks, tape of any kind, floppy disks, older hard disks, some CD formats) are a matter of urgency and must be transcoded in the very near future.

Much of the data that anthropologists collect is not suited for representation in simple flat files; indications of complex contexts are as much a part of a dataset with respect to description and analysis as the data items themselves, and complex structures and relationships cannot be easily represented as a simple flat "row and column" type database.

So data organization, and related metadata, is often multidimensional, taking forms that can be represented in trees, graphs or other relational abstractions. Additionally, metadata is used to relate how higher order classifications, inferences or transient references might be represented and processed. Semantic services are usually built on new layer of metadata applied to a "flatter" syntactic metadata layer.

Semantic interoperability for datasets is not as straightforward as syntactic interoperability. Whereas syntactic conventions allow us to differentiate data items and recover metadata relating to these, semantic operations are required to identify similar data items between datasets. One traditional method uses a codebook describing each variable and its possible instantiations in the data set. Similarly, the most common means of supporting semantic interoperability between digital datasets is through associating metadata with each data item and value. Metadata is usually (but not always, as for XML) maintained apart from the data itself, serving as a kind of template for a class of data sources, rather than a descriptor for a single dataset or individual data items. Most people are familiar with simple metadata, such as that for a publication in bibliographic record, where slots are specifically designated for particular roles or states; author, title etc. Metadata includes syntactic information regarding how data will be organized in addition to providing a label for a slot. The goal for semantic interoperability is to relate the items between datasets to link similar items. For interoperability, information must also be available relating the possible values for data and how these relate to each other. For example, for the simple variable ‘age’ one dataset might relate this in years, another years and months, and a third corresponding to the set {child, adolescent, adult, elderly}. With conversion, the latter will clearly not be equivalent to the former, just comparable.

In the case of ethnographic data, such as that processed by HRAF, while there are often tables of synthesized aggregate data, the majority of the data is natural language text (the majority in English) organized into conventional structures associated with published material. Syntactically, this makes it fairly easy to put a range of material from different datasets (individual documents) into a common form, particularly with sufficient metadata relating to elements of publications, and a relatively robust XML schema that can represent a wide range of publication formats.

Although some of the publication metadata has a semantic definition, relating time, place, and society, the substantial semantic metadata HRAF has added are the Outline of Cultural

Materials (OCM) subject classifications applied by anthropologically-trained HRAF analysts. These allow, for example in searching for information about a particular topic such as how marriages are arranged using the OCM code "Arranging a Marriage" in an Advanced Search which ensures that the text of a given search unit contains information relating to the topic. However, in the current version of eHRAF, the researcher has to make his or her own decisions about the forms of the arrangements. Are arranged marriages customary? If so, how are they arranged? Or, do individuals decide on whom to marry? If so, are there customary patterns of courtship? In other words, after finding the passages with information, researchers are then on their own.

Pragmatic interoperability recognizes that for standards to be adopted by the anthropological community these must have clear benefits, have useful levels of partial adoption, and be researcher-extensible and open to application for legacy data sets. Identifying the pragmatic requirements for data re-use in interdisciplinary research goes beyond simply matching up elements syntactically, structurally and semantically. For example, relating the interaction of ethnographic research with genetic research is not a simple matter of identifying and structuring data, but requires some form of common reference or context, such as applying to a common population or site.

Steps HRAF Is Taking To Improve Interoperability

In HRAF development we are using several approaches to promote syntactic interoperability. The most important change dates to the first online version of the database when SGML was used to mark up the text, structure of the text, and associated metadata, which was then converted to XML in the mid-2000s and published in XML in 2008. XML provides the following advantages: 1) it is far easier to perform a range of transformations of the document text and context; 2) it makes a wider range of queries much easier to perform, including queries that reference the context of matching entries; and 3) it promotes transformation into more specialized outputs that can be further processed for further analysis and customized reporting. Even though each document has a unique structure and content, these are syntactically encoded using the same XML schema, and makes interoperability from these heterogeneous sources at least possible.

To increase semantic interoperability in the future, we plan to base further services on searching a transformation of the present HRAF Production XML schema into what we call

the HRAF vDoc XML schema (vDoc stands for “virtual document”). In contrast to the Production schema, which tries to reproduce the structure of the original documents, the vDoc schema reduces some of the heterogeneity between works by standardising relevant document structure and context, making relevant metadata available at the level of an SRE or search and retrieval unit (usually a paragraph) instead of having to retrieve it from the document context. We also plan to add multiple versions of the text in each record including the original markup, a text only version, a record of the parts of speech for the plain text content. We will also add statistics relating to the text, ranging from simple frequencies within the search unit, and an indication of which terms have ‘gravity’ (also appear in prior and/or subsequent search units). There will be other additions to support semantic interoperability. The vDoc XML structure will provide a uniform way to represent both the base data and references to the base data. Within the HRAF services framework each service will produce a vDoc as output, excepting a few whose purpose is to render vDocs into forms for display or interchange with other tools and platforms. However, most services transform or aggregate other vDocs into a new composite vDoc. vDocs are flexible enough to embed most other data formats, so can serve as an all-purpose media for compositing different data streams and types.

The use of topic-mapping in conjunction with OCM classifiers will be at the intersection of semantic and pragmatic interoperability as it will leverage the OCM and the decisions analysts make with respect to the OCM in conjunction with topics emerging from the texts, hopefully reflecting the ethnographers’ intentions. Pragmatic interoperability supports the researcher’s capacity to answer specific questions while drawing across data from different sources. The OCMs alone are a powerful tool in this respect, and the basis for the degree of success the HRAF databases enjoy. But the extent of manual labor required to utilize the material once located is quite onerous. In its most basic use topic-mapping will help reduce effort by helping the researcher to reduce the set of results examined, since topic-mapping provides much more detail regarding the contents than the OCM classification.

Pragmatic interoperability is primarily about integrating across platforms and disparate models. We will turn to this next in the context of a plan for a service platform to incorporate new ethnographic data from others.

Towards a Service Platform for Broader Re-Use

Beyond the present HRAF resources, we are working towards expanding access to and reuse of other researchers' underlying and published ethnographic and other data, without compromising confidentiality or other constraints, to promote reuse of data generally in a new services platform. There are considerable problems with publishing most of the material that an ethnographer collects during fieldwork in a given society. Most of the information is highly personal and often sensitive, and agreements for use often involve per person and occasion agreements,. Even then, the ethnographer has a duty of care that goes beyond the agreement of the people involved. The topics researched often contain highly confidential material that has a high potential for personal, political, or even legal damage. Because the record is cumulative and detailed, even redaction and anonymizing is inadequate to protect interests, as it is often quite simple to work out the identity of specific individuals.

One partial solution to the confidentiality problem is to search within underlying text, other than for embargoed terms (such as personal names and place names) and topics (such as potentially subversive activities) designated by the contributing researcher, but to not return the matching underlying text. Rather, a range of transformations of the results are made available to be reported (as permitted by the contributor). These include basic information, such as word frequencies for the results or sub-units of the results and also the embedding document context, but also topic maps for the results, and various approaches to narrative or more structured summaries of the content. Further leveraging the HRAF collection, examples from HRAF similar to those for the undisclosed text results would be available so that researchers can see the range of material across a designated range of HRAF cultures that corresponds to the themes in the confidential material. In addition to their exemplary value, these will assist in interactive auto-coding of the unseen material.

Drawing on correlations between HRAF derived text which includes OCM classifiers, and texts submitted by other researchers which do not, a graph can be created with will facilitate assignment of OCM codes to the external texts, either automatically, or with some assistance by the depositing researcher. A services platform for detailed topical analysis of underlying texts together with outer textual metrics will be configurable to different topical graphs and, potentially, ontologies other than the OCM. Tools will be developed to assist researchers in situating their own results with respect to others for comparative purposes using material submitted by other researchers in addition to the HRAF materials. These

services will operate on sources directly submitted by researchers to the repository and published material, for which underlying text will not be quoted, and additionally the expertly analyzed, classified and curated ethnographic data corresponding to a number of accepted samples (the HRAF Collection of Ethnography) using standardized topical tags from the Outline of Cultural Materials.

Some of the material we will want to integrate with the HRAF corpus will have been developed with other schemas and models. To achieve pragmatic interoperability, we will be exploring a method analogous to the notion of "docking" originally proposed by Axtell et al. (1996) for agent based modelling. Docking is a method of establishing connections between apparently disparate models which have some matching or related classes, variables or parameters. Collaborating researchers find common ground where their respective descriptive understanding of their datasets (perhaps as a model) are sensitive directly or indirectly to each other, collectively building a new layer that "glues" the two together. However detailed knowledge of the others' model is not necessary for a given researcher, only the agreed model that connects the two. Pragmatically, a "docking" approach helps researchers focus mainly on their own research problems but allows them to further explore the impact of the larger context on their research problem. For example, by relating marriage, the distribution of a relatively rare allele and mortality the social anthropologist, biological anthropologist and demographer each gain insight into how context refines more general understanding.

Conclusion: Expanding HRAF Research Services

Currently the eHRAF application available to the membership has a fixed set of options for search and reporting, fairly typical of current generation web applications. But we are repurposing search and retrieval operations as a variety of services that are independent of any specific web application. 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 will retain this structure for production and archival purposes, but to facilitate large scale search and retrieval services we are normalising 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. Our

principal goal is to develop new tools and infrastructure to support primary and secondary ethnographic research using the data resources available at HRAF.

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 are working on 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 work towards auto-coding techniques to promote broad consideration of comparative analysis and situation of human practices and behaviors for basic and applied research. To support data mining we are developing services with which we are experimenting with different approaches to producing 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 are also exploring methods to apply similar auto-classification to other sources, ranging from academic publications in anthropology to newspaper articles. Finally, beyond the present HRAF resources, we are working towards expanding access to and reuse of other researchers' 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.

References

Divale, William T. 1977. Living Floor Area and Marital Residence: A Replication. Behavior Science Research 26 (2): 109–15.

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. and Melvin Ember. 2009. Cross-Cultural Research Methods, 2nd edition. (Lanham: AltaMira)

Ember, Melvin. 1973. An Archaeological Indicator of Matrilocal Versus Patrilocal Residence. American Antiquity 38 (2): 177–82.

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.

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

Murdock, George Peter, Clellan S. Ford, Alfred E. Hudson, Raymond Kennedy, Leo W. Simmons, and John W.M. Whiting (New Haven, Conn.: Human Relations Area Files).

Porčić, Marko. 2010. "House Floor Area as a Correlate of Marital Residence Pattern: A Logistic Regression Approach." Cross-Cultural Research 44 (4): 405–24.

Porčić, Marko. 2012. "Effects of Residential Mobility on the Ratio of Average House Floor Area to Average Household Size: Implications for Demographic Reconstructions in Archaeology." Cross-Cultural Research 46 (1): 72–86.