Available online at www.sciencedirect.com

SciVerse ScienceDirect

Poetics 41 (2013) 366–383

POETICS

www.elsevier.com/locate/poetic

Rapid ethnographic assessment for cultural mapping

Tracy Van Holt [a,*], Jeffrey C. Johnson [b], Kathleen M. Carley [c], James Brinkley [d], Jana Diesner [e]

[a] *Institute for Coastal Science and Policy, Geography Department, East Carolina University, Greenville, NC 27858, USA*

[b] *Institute for Coastal Science and Policy, Sociology Department, East Carolina University, Greenville, NC 27858, USA*

[c] *Institute for Software Research, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA*

[d] *Institute for Coastal Science and Policy, East Carolina University, Greenville, NC 27858, USA*

[e] *The iSchool (Graduate School of Library and Information Science), University of Illinois at Urbana-Champaign, Champaign, IL 61820-6211, USA*

Available online

Abstract

Today researchers need an efficient and valid approach to mine and analyze the large amount of textual information that is available. Automated coding approaches offer promise but a major concern is the accuracy of such codes in capturing the meaning and intent of the original texts. We compare the recall (number of codes identified) and precision (accuracy of the codes) that included bodies of texts coded (1) manually by humans based on the Outline of Cultural Materials (OCM) code book, (2) semi-automatically by computers that used a human-generated content dictionary containing Rapid Ethnographic Retrieval (RER) codes, and (3) automatically by computers that used an automated version of the OCM content dictionary (AOCM). We applied network visualization and statistics to quantify the relative importance of codes. The semi-automatic coding approach had the highest balance of recall and precision. Network visualization and metrics identified relationships among concepts and frame codes within a context. Semi-automated approaches can code much more data in a shorter period of time than humans and researchers can more easily refine content dictionaries and analyses to address errors, which makes semi-automated coding a promising method to analyze the ever-expanding amount of textual information that is available today.

Keywords: Content analysis; Accuracy; Network analysis; Data mining

* Corresponding author.

0304-422X/$ – see front matter
http://dx.doi.org/10.1016/j.poetic.2013.05.004

1. Introduction

There has been an explosion in the number of available digitized textual sources such as newspapers, journals, blogs, the (participatory) web, etc. Given the sheer magnitude of textual data, it is not possible for human coders to keep up with the flow of text-based information. This article is a first step toward understanding the strengths and weaknesses of semi-automated coding of texts by assessing the accuracy of these coding methods, particularly dictionaries. We seek a coding approach that has the highest balance of recall (number of concepts coded) and precision (codes the concept correctly). In addition, we visualize and analyze word co-occurrences using a network approach providing opportunities to quantify the relationships among codes. If semi-automated coding of textual data and network analysis can mimic human coders' ability to code for ethnographic concepts, then the door opens to new types of ethnological, comparative or cross cultural studies of relationships at multiple spatial and temporal scales previously unheard of.

This matters because textual data available online via newspapers, Twitter, blogs, etc. provide rich sources of data by which to model various aspects of human behaviors—all the way from individuals to societies (NRC, 2008). For example, Internet applications are moving toward technologies that facilitate interaction and participation with the end user, which offer insights into human behavior (O'Reilly, 2005). If people's interactions, such as a comment on Twitter, are geospatially tagged, or if locations in news articles are geospatially coded, then we can code for the patterns of human opinion and behavior can be mapped across space and time (Rinner et al., 2008; Van Holt et al., 2012). Texts provide fine-resolution data for events that are rare, such as conflicts (Mack, 2007), and therefore difficult to forecast because of a lack of sufficient data (Schneider et al., 2011). Data used in forecasting models are often at the macro level and, as a consequence, researchers may not pick up on finer resolution indicators that may help improve accuracy (Schneider et al., 2011). However, researchers such as Brandt et al. (2011) have advocated using online data resources to forecast conflict in real time. By rapidly and systematically analyzing text-based data, we can analyze human perception and behavior through time and move toward forecasting events and the likelihood of scenarios. Of course, the coding of opinion patterns, the mapping of human behavior, and the forecasting of events require a useful approach.

1.1. Rapid ethnographic assessment

We are developing a computational system that will help to advance rapid ethnographic assessment. We make a distinction between ethnographic and other kinds of assessments in that ethnographic assessments follow in a tradition of the coding and analysis of ethnographic texts using abstract codes that capture important theoretical constructs (e.g., kinship lineage systems) and that attempt to maximize the contextual properties of the original texts. Rapidly analyzing a culture, the socio-economic and environmental drivers of culture, and how these processes change over time all require a systematic and robust way to extract and analyze data. This system (see Fig. 1) will operate end-to-end so as to enable the researcher to rapidly collect and relate data across a broad range of situations; to understand and identify fundamental dynamic processes; and to feed ethnographic data into both qualitative and quantitative models. One such example is from Van Holt et al. (2012), who modeled ethnic conflict and peace in Sudan and South Sudan using newspaper articles. They concluded that ethic conflict was associated with livestock, environmental resources, and the structure of multi-ethnic group ties to these resources: ethnic

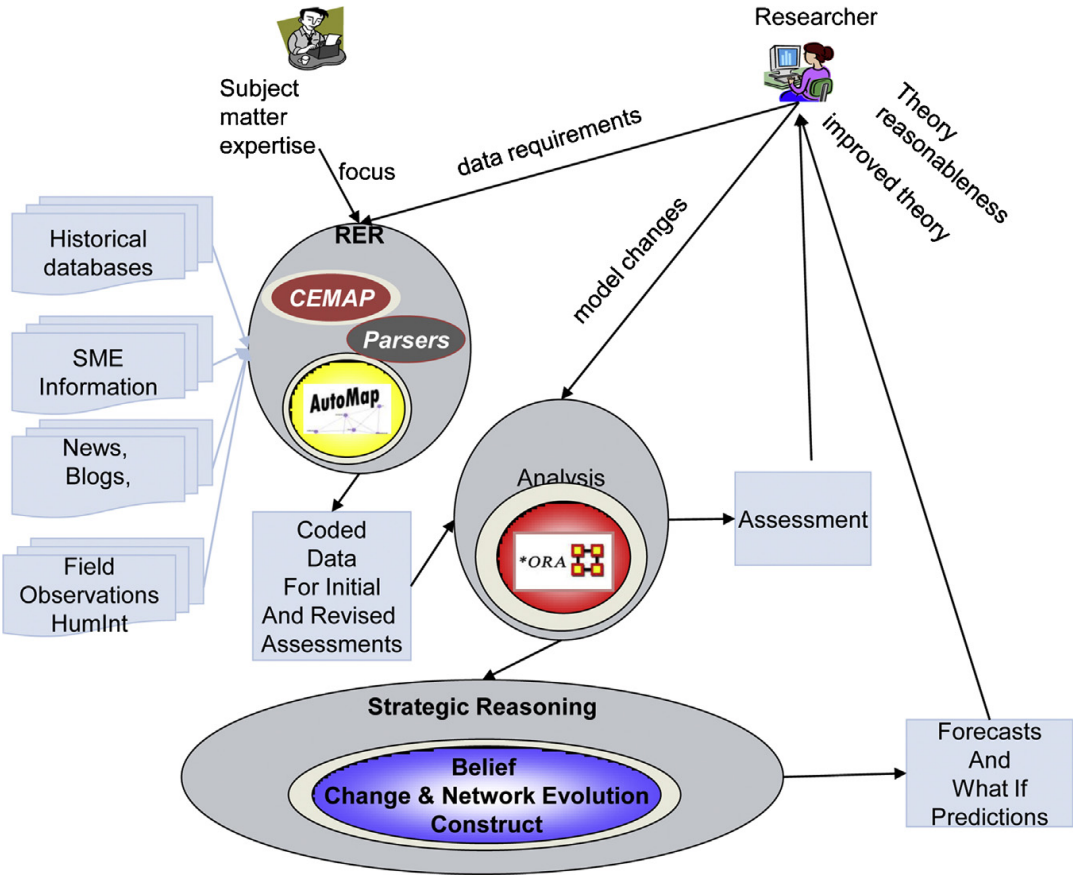

Fig. 1. Overview of system.

groups associated with peace had more ties to the coded concept ‘‘biomes’’ (forest, rivers, etc.). We believe such a system will fundamentally alter the way in which ethnographers, ethnologists and modelers work together and support improved data collection, model building, accuracy assessments, and model validation. However, going from text data to various kinds of models (e.g., network models) involves multiple points for introducing and propagating errors (Diesner, 2012). Therefore, key components in this process are to make sure that we understand the errors in coding and how these errors may affect our ability to understand and construct valid models of human behavior.

This article is a first step toward understanding the strengths and weaknesses of semi-automated coding of text-based material; it does so by assessing the accuracy of the coding process, especially of coding dictionaries, which are a key component of the presented approach. It is referred to as ‘‘ethnographic’’ in that we follow in the tradition of cross-cultural researchers, such as George Peter Murdock, in creating standardized codes that have ethnographic meaning capturing elements of culture, society, and economics. We compare two automated coding approaches against a gold standard—namely the Human Relation Area Files (HRAF), which are documents (i.e., ethnographies) that were manually coded by researchers using the Outline of Cultural Materials (OCM) code book, one created over the span of George Peter Murdock’s career. In comparison, the Rapid Ethnographic Retrieval (RER) represents the semi-automated approach that includes a dictionary, which took a year for people to develop, and software (AutoMap and ORA; Carley et al., 2011a,b, respectively) that automatically coded and analyzed the data based on a dictionary that involved humans in the analytical loop. Although somewhat labor intensive, the value of these two dictionaries is that they can be modified for use in the

analysis of other texts with little effort. In contrast, the automated approach used a dictionary that was generated in one day by automating the OCM codes (AOCM) and software (AutoMap and ORA) that automatically coded and analyzed the data with little human involvement. We compare the three coding approaches in terms of recall (number of concepts coded) and precision (codes the concept correctly).

1.2. Ethnographies and ethnologies

There is a long history of using coded texts to produce large-scale models of human systems that include ethnographies—studies of individual cultures and ethnologies, or cross cultural comparisons. Ethnographies usually require extensive field work to gather such descriptions. Researchers often try to understand the lived experience (Spradley, 1979) of informants through ethnographies, which can be extracted from narratives—such as open ended and structured interviews, as well as historical documents—through participant observation. In one tradition, after collection, a large corpus of these rich narratives are then coded. However, these coding endeavors have been very labor intensive—involving a large number of coders, usually graduate students, who would code vast volumes of ethnographic materials with respect to a predetermined set of themes, in part to restrict the work to a manageable workload. Most notable advances in ethnologies are based on ethnographies from multiple researchers that were classified by George Peter Murdock who produced several cross cultural data sets (Whiting, 1986)—including the Standard Cross-Cultural Sample—that classified distinct cultures and societies as the units of analysis for cross-cultural studies and a large number of ethnographically coded variables. These data sets were derived from the corpus of known ethnographies of the time that were collected through extensive field work by social scientists and the coded variables were produced by human coders—mostly coders not a part of the original ethnographies. The coding of the texts was labor intensive, but their efforts were able to produce reliable codes for often abstract concepts and variables across a large sample of societies from around the world. Such coded data could be used to model, study and account for patterns observed across cultures that might help to illuminate, for example, the cultural factors related to animal husbandry or matrilineal descent. This coding enterprise took place in the 1950s and 1960s, at time when the number of ethnographic texts, although growing, were still relatively limited in number.

Today, ethnographies are evolving from one-shot case studies focused solely on the lived experience in a cultural setting (Spradley, 1979) to include studies that integrate measures of social structure, as well as multi-disciplinary, multi-temporal (Abello et al., 2012), and multi-group analyses. For example, Johnson and Orbach (2002) have found that the structural features of the social system, such as social networks, show how personal relationships are related to success in establishing political change. Experiential and environmental factors are increasingly viewed as important drivers that shape culture or shared knowledge (Boster and Johnson, 1989; Reyes-Garcia et al., 2004). Recent behavioral change or adaptation studies include geospatial and ecological factors, in addition to traditional measures of knowledge and technology (Van Holt, 2012). Still longitudinal studies are rare, despite the fact that this type of experimental or analytical design provides the strongest scientific evidence to understand changes in human behavior. One such compelling study is Godoy et al. (2005), which tracks informants' knowledge through time as they became integrated into the global market. As studies discover more factors that are predictive of socio-cultural behavior, ethnographic studies become increasingly interdisciplinary requiring larger teams, more time and greater expenses to cover this expanding set of factors. There also may be higher risks and costs in research conducted in regions of high

conflict. For all these reasons, a more scalable approach to collect ethnographic information is needed, and digital texts offer this opportunity.

1.3. Network analysis of word co-occurrences

Once coded, the themes can then be analyzed by word frequency, word co-occurrences, and network analysis (Bernard and Ryan, 2009). The relationships among themes can be displayed and further analyzed via multidimensional scaling, clustering algorithms, network analysis and visualization techniques such as spring embedders (Bernard and Ryan, 2009; Johnson and Krempel, 2004), or they can be geospatially mapped (Van Holt et al., 2012). Word co-occurrences characterize relationships and interactions among actors and events, behavior, etc. Osgood's (1959) analysis of the co-occurrence of words in W.J. Cameron talks on the *Ford Sunday Evening Hour* radio program was the precursor to the use of network analytics to study and model texts. Osgood coded the talks (37 in total) according to 27 concepts; tested the significance of the associations; and displayed the synthesis of the topics in a crude multi-dimensional scaling visualization, where the size of the node (word) reflected its prominence and ties (arcs) signified that two concepts co-occurred during a given talk. Carley (1994) has shown that a map analysis of texts (also a precursor to network analysis) is a way to integrate cognition and culture; one example in the study discusses how the definition of robots has changed over time from being perceived in a negative to a positive light. Center resonance analysis (CRA), developed by Steve Corman and Kevin Dooley, uses network analytics to produce abstract representations of text by linking together words in texts (CRA, 2001). However, CRA is a data reduction method that involves a combination of natural language processing and network methods to produce an abstract representation of the overall content in the original text(s).

Moving away from network analysis, Bourdieu and Wacquant (1992) argued that correspondence analysis is useful for analyzing relational thinking, specifically that of field theory. Nonetheless, de Nooy (2003) argued that network analysis can provide additional information that correspondence analysis cannot, which is a way to quantify social capital and other person to person relationships. Johnson and Griffith (1998) confirmed some of these same findings but advocated the integration of the two approaches (correspondence analysis and network analysis) as both are relationally based but provide different types of information about the nature of such relations. That said, Krinsky (2010) coded news items from a LEXIS-NEXIS search on ''Workfare and New York City'' for actors (judges, state officials, advocates, etc.) and their political claims. Network analysis was used to identify conversations and bridging power across conversations of actors that spanned across multiple claims. Thus, a network approach to analyze relationships offers potential to provide more context to the automated content approach, increasing the ability to mimic human coders.

1.4. Semi-automated text analysis

Content analysis is a methodology for summarizing and assessing the content of texts (Holsti, 1969; Neuendorf, 2002). This approach mainly relies on content dictionaries (i.e., tables of text terms), which can consist of one or multiple words that are associated with a particular theme (Carley, 1994; Gerner et al., 1994; Roberts, 1997; van Cuilenburg et al., 1986). Content dictionaries used to be developed by hand, but automated approaches have been developed, applied and evaluated to accelerate this process (Diesner and Carley, 2008; Diesner, 2012; King and Lowe, 2003; Schrodt and Gerner, 1994). Most of these approaches use a (mixture of) lexical, syntactic,

semantic, logical and statistical information from the text documents and sometimes meta-data on the texts' data, such as key words and index terms (for a review, see Diesner and Carley, 2010).

Automated text analysis began in the 1960s by Philip Stone, who developed the first dictionary-based content analysis program, the General Inquirer (Stone et al., 1962). Today the General Inquirer is associated with some of the most developed code books for coding text. These include the Harvard IV-4 dictionary that contains 182 categories. The dictionary also has valence categories that are used to refine the meaning of the text. Coding for valence has recently started to gain momentum again under the labels of sentiment analysis and opinion mining. Additional code books have been added, although it is unclear how the legacy of the Harvard IV will evolve. The Kansas Event Data System (KEDS), including the Textual Analysis by Augmented Replacement Instructions (TABARI), is also a dictionary-based approach that codes political events into predefined and continuously updated categories (Gerner et al., 1994). Most content analysis software supports the construction of dictionaries, but it may be cumbersome to consistently and coherently customize and integrate multiple dictionaries for analytical efforts other than for what they were originally intended.

Our approach combines (1) semi-automated content analysis, (2) mapping text terms to concepts (e.g., ''Omar Hassan Ahmad Al-Bashir'' and ''Bashir'' both to ''al Bashir''), (3) mapping concepts to ontological categories (e.g., ''al Bashir'' to ''agent''), and (4) a link extraction technique (Carley et al., 2007; Diesner and Carley, 2008). In steps one and two, a dictionary is used that is seeded with a set of generalizations and aliases, as well as automated solutions for stemming and reference resolution, to identify a set of concepts. Using a human-in-the-loop approach referred to as data-to-model process (D2M), the human researcher can apply further generalizations and identify aliases to reduce the concept set (Carley et al., 2012). For example, a trained coder reviews the results of automated strategies and has the chance to make changes to automatically generated outputs and solutions. The result from this phase of the coding, like other content analysis procedures, is a list of concepts and their cumulative frequency, which is used to improve and expand a dictionary. In step three, we classify concepts into a set of ontological categories. These categories are agents (e.g., people), organizations (e.g., tribes), locations (e.g., New York City), tasks (e.g., conflict), resources (e.g., potable water), and others. This is accomplished using a probabilistic entity extraction model trained via Conditional Random Fields; a supervised machine learning technique appropriate for working with sparse, large-scale data (Diesner, 2012; Diesner and Carley, 2008). Then, using a human-in-the-loop approach, the suggested classification is vetted—resulting in improved coding (Carley et al., 2012; on the impact of this verification step, see Diesner, 2012). The fourth step of the coding process is the extraction of links among instances of ontological classes. This is done automatically using a proximity-based approach, where the user specifies a window size within which all concepts are linked to each other (Danowski, 1993; for the impact of proximity-based coding and related error rates, see Diesner, 2012). The result of the overall D2M process is a network representation of the textual information, that is, a network where each concept is associated with an ontological category. The extracted concept networks contain weighted, bi-directional links. The weight indicates the cumulative frequency with which a link was observed. The coded text is then analyzed and visualized with the ORA software (Carley et al., 2011b).

1.5. Illustration of going from texts to networks

Our dictionary codes for different ontology categories, such as knowledge, resource, task, etc. For a link to appear in the network, the coded concept appeared within *n* words of another coded

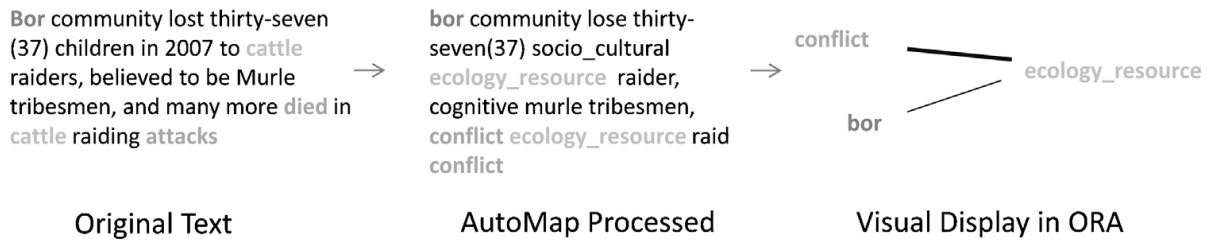

Fig. 2. An example of how a text is coded as a network.

concept, where *n* is set by the user. The concept networks can be viewed per ontological category or they can be integrated. In this example (Fig. 2) text from the *Sudan Tribune* was extracted and a user-generated content dictionary was applied by AutoMap. After the processing, **Bor** would code to *bor*, an ethnic group in Sudan, and then cross-classified as an organization. **Cattle** would code to *ecology_resource* and then classified as a resource; and the word **attacks** codes to *conflict* and then classified as a task. To generate the networks, we set a window size of seven, which means that coded text within seven words of each other show a link. This window size was found to be optimal for our data set, and it resulted in not too many or too few links in the code networks. AutoMap generates an xml file (dynetml) that is imported into ORA, a network analysis software system that reads each ontology category (i.e., knowledge, resource, and task, etc.) as a separate network and also integrates ontology categories. In ORA, social network metrics can be computed. Moreover, network visualization can be generated that show the structure, patterns and strength of relationships among categories and the network overall. For example, in Fig. 2, we see that the tie between *conflict* and *ecology_resource* is thicker than other codes—indicating multiple co-occurrences of these concepts in the text and a strong link between conflict and ecology-resource, in this case, cattle. One concern, however, is how accurate are these codes in capturing the meaning and intent of the original texts? In investigating this issue, we now turn to the accuracy of our content dictionaries in comparison to other methods for coding including the use of human coders only and an automated approach. In addition, we demonstrate the visual and analytical capabilities of using AutoMap and ORA in the context of coding ethnographic data to answer substantive questions about social groups.

2. Research design and methods

Here we analyzed 25 paragraphs of an HRAF coded ethnography, *Shaping of Somali Society: Reconstructing the History of a Pastoral People, 1600–1900* (Cassanelli, 1982), where human coders hand-coded each paragraph according to the OCM code book. We compare the human-coded OCM data to computer processed data coded by the automated content dictionary of the OCM categories, which we will call AOCM categories, and the semi-automated Rapid Ethnographic Retrieval process (RERs).

The OCM coded files were obtained from the HRAF Archive of Ethnography that includes over a million pages of ethnographies, where each paragraph has been hand coded or more precisely indexed by professional anthropologists according to culture traits and cultural groups using the OCM categories (Murdock, 1987) and Outline of World Cultures (OWC) (Murdock, 1983). The intent of coding each section (a single paragraph or group of paragraphs) in the ethnographies is to facilitate cross-cultural research (ethnological studies). Researchers identify a cultural trait of interest and sample texts—in this case, ethnographies—containing those traits

across multiple cultures using the Standardized Cross Cultural Sample (Murdock and White, 1969) or other sampling techniques. The HRAF are electronically available and the OCM categories span a wide range of culture traits, thus making HRAF an ideal source of texts to compare coding approaches. The OCM categories are posted at the end of each paragraph or group of paragraphs that have been manually coded by HRAF employees.

To generate the automated AOCM codes, we used a highly efficient approach. We spent one day automating the coding of the OCM categories. First we downloaded the descriptions of each of the OCM categories as provided on the HRAF website or in the OCM manual (Murdock, 1987). We then parsed out each independent word or phrase using punctuation characters and created a dictionary, also known as thesaurus, where each independent word or phrase mapped to the corresponding OCM code, which we then call an AOCM code. The OCM descriptions are incomplete; for example, the fishing category (OCM_226), lists fish and shellfish in the description, but does not include algae. In other cases, the descriptions are difficult to automate; for example, economic importance of fishing, a description OCM_226, was simply not mapped for. The purpose of our comparison is not to automate each OCM code precisely, but rather to see how by searching for a few of the descriptor words provided in the OCM guidebook, fully automated techniques compare to manual and semi-automated techniques. Even if only a small fraction of the text is coded correctly, we hope to understand what the computers code efficiently, what human-coders are required for, and how to improve upon this fully automated coding approach.

In contrast to this efficient technique, it took us over a year to create the RER categories by providing lists of words within each category that helped create our dictionary. Each term in the data set goes through a hierarchical classification system following from the term itself to the RER code. The classification system begins with the broadest categories (i.e., the ontological categories that include knowledge, resources and tasks, etc.). All of the verbs and their variants that the human coders observed or could think of were assigned to the categories in the task ontology. All terms that are resources, such as oil, cattle, or diamonds, were assigned to their specific categories within the resource category. Other complex concepts representing information were assigned to categories within the knowledge class. Each of these categories has several sub-categories called RERs for Rapid Ethnographic Retrieval. The analyses presented here use the finest level of resolution of RER categories. The RER approach is not a direct match to the OCM, but we did use this code book as a guide to make sure that we characterized the majority of cultural materials. At this time, a word can only code to one category, though in the future, we plan to add multiple categories.

We compare the total number, as well as the content of codes found (recall) and accuracy of the AOCM, OCM, and RER approaches (precision). We then compare the percentage match from the human coded texts (OCM) and automated coding approaches (AOCM and RER). We identify what types of additional codes the automated approaches are picking up on (false positives). We then compare the automated approaches (AOCM and RER) for coding frequency and accuracy for each coding event. Finally we visualize the AOCM and RER networks, and compare the relationships among codes using network statistics.

The RER network contains coded concepts from three different ontologies—knowledge, task and resource—that are each represented in the network with a different icon. These categories were selected as they are central to our content domain. The AOCM does not have these distinct categories, only one network is generated, and all nodes are represented with an icon. In order to make the two networks more comparable to each other, ORA was used to merge the three RER ontologies into one. The task and resource ontologies were moved into

knowledge. This allows for network level metrics to include all connected nodes in order to produce a single set of measurements. For visual clarity, only the main component of each network is visualized and all isolates (i.e., nodes not connected to any other node) were deleted. Both networks were symmetrized and then characterized by computing the degree centrality and betweenness centrality for each coded concept. Degree centrality measures how many times a concept was adjacent any other concept; a concept with high degree is tied to many other concepts in the text. The betweenness centrality metric is a measure of the extent to which a code or node is on the shortest path between all other nodes; it reflects how important a node is in connecting other nodes or bridging concepts. Since the OCM data were coded at the paragraph level they were not included in the network-level comparison to avoid conflicts of comparability.

3. Results

3.1. Accuracy assessment and unique codes

The highly efficient automated AOCM technique had a high recall rate but low precision, i.e., it identified the highest number of unique codes (138); however, this approach was also the least accurate[1] (only 38% correct matches; see Table 1). The human coded (OCM) technique resulted in low recall but high precision in that it found the fewest codes, but these codes were all accurate (100%). The content dictionary based semi-automated RER approach was also highly accurate (96%) and identified 84 codes, and it had the highest balance of recall and precision out of the tested methods.

Out of the 33 different concepts coded for in the human-coded OCM, twenty-four (73%) were picked up by the AOCM automated approach (Table 2). Twenty-nine of the OCM codes (88%) and 42 (80%) of the AOCM codes could be accounted for in the semi-automated RER approach through similar code matches (similar matches were used because the RER code book does not precisely match the OCM or AOCM code book). The RER approach correctly coded for 24 additional concepts not picked up by the AOCM and 46 additional codes not picked up by the OCM.

In comparing the automated and semi-automated approaches (all coding events), the automated AOCM approach coded (or recalled) 593 items of text throughout the 25 segments analyzed, of which 32% of the coding events were accurate (precise) because they reflected the context of the original text (Table 3). The semi-automated RER coded (recalled) many more items of text (824) than the AOCM, and nearly tripled the accuracy (precision) with 93% correctly coding for the concept.

3.2. Example of errors and breadth of coding

The following texts are excerpts from *Shaping of Somali Society: Reconstructing the History of a Pastoral People, 1600–1900* (Cassanelli, 1982), and they demonstrate coding with OCM, AOCM, and RER codes at the sentence level. In the following texts, the bold words coded to the italicized terms in parentheses. The codes are differentiated in terms of AOCM or RER classification. The OCM codes are at the end of the text.

[1] A code was considered accurate if it reflected the text at least one time.

Table 1
The semi-automated RER approach has the highest balance between recall (total codes identified) and precision (codes that were accurately coded). The automated AOCM approach has high recall but low precision. The human coded OCM codes had the highest precision but lowest recall. Accuracy was determined by a human coder.

Coding method	Recall	Precision			
	Total codes found	Accurate codes		Inaccurate codes	
		Total	Percent	Total	Percent
OCM	33	33	100%	0	0%
AOCM	138	52	38%	86	62%
RER	84	81	96%	3	4%

Note: Accuracy was determined if the code was correct at least 1 time.

Table 2
Percent match among human coded (OCM), automated (AOCM) and semi-automated RER coded concepts. Number of additional codes found in each comparison identified as well.

	AOCM			RER		
	Number match	Percent match	Total additional codes found	Number match	Percent match	Total additional correct codes found
OCM	24[a]	73%	28	29[b]	88%	46
AOCM				42[c]	80%	24

Note: Accuracy was determined if the code was correct at least 1 time. Total additional codes found in the RER analysis may not sum to 81 because multiple RERs may closely match a single OCM or AOCM code.

[a] See Appendix 1 for a list of OCM and AOCM codes found and those that matched.

[b] See Appendix 2 for a list of OCM and RER codes found and those that matched.

[c] See Appendix 3 for a list of AOCM and RER codes found and those that matched.

Table 3
In terms of total coding events—individual times a concept is coded, the semi-automated coding (RER) had the higher recall and precision in comparison to the automated (AOCM) approach.

ACOM		RER	
Recall (total coding events)	Precision (percent of events that were accurate)	Recall (total coding events)	Precision (percent of events that were accurate)
593		824	
190	32%	765	93%

Although they were **outstanding** *(RER adjective)* breeders of **camels** *(RER livestock)*, **cattle** *(AOCM 233 pastoral activities; RER livestock)*, and **horses** *(RER livestock)*, they never raised **animals** *(AOCM 363 streets and traffic; RER fauna)* **specifically** *(RER adverb neg)* to meet the needs of any large nonpastoral **population** *(RER population).*

In the northern part of the country, these goods were available annually at the winter **bazaars** *(AOCM 443 retail marketing)* in the **coastal** *(AOCM 131 location; RER biomes and land cover)* **towns** *(AOCM 632 towns)* and periodically from itinerant traders who set up makeshift shelters near **water** *(AOCM 318 environment quality; RER biomes and land cover)* holes or wadi **s** *(AOCM 527 rest days and holidays)* in the **interior** *(AOCM 131 location; RER valence).*

OCM codes: *221_annual_cycle, 233_pastoral_activities*, and *439_external trade.*

All coding schemes picked up on pastoral activities well. An example of a positive match is *AOCM 233 pastoral activities* and *RER livestock* both coding for **cattle**. **Water** was coded as *AOCM 318 environment quality* under the AOCM code book, which is in the same domain, but not a precise code. The OCM codes picked up on an *OCM_221 annual cycle*, which neither the RER nor AOCM codes accounted for. The RER in addition, coded for parts of speech, population, valence, and biomes and land cover. False positives in the above text include *AOCM 363 streets and traffic* for **animals** and *AOCM 527 rest days and holidays* for **s**.

> Gellner formulates the distinction as ''simple'' vs. ''symbiotic'' **nomadism** *(AOCM 221 annual cycle; RER livestock)* and he sees the **latter** *(RER adjective)* as **characteristic** *(RER valence)* of **Middle** *(AOCM 911 chronologies and culture sequences)* Eastern **pastoralists** *(RER livestock).*
> OCM codes: *221 annual cycle, 233 pastoral activities, 619 tribe and nation*

The AOCM and OCM coding of the above text both coded for *annual cycle*. The ACOM picked up on the word nomadism, which was an element of the OCM dictionary, whereas the OCM coder read the paragraph and coded the paragraph as *annual cycle*. The RER and OCM coding both picked up on *livestock*. Only the OCM coded for *tribe* and *nation*. The AOCM approach falsely coded *AOCM 911chronologies and culture sequences* for **Middle**.

3.3. Network-level analysis

The network-level characteristics generated via the semi-automated RER and automated AOCM show that much more information is generated via the RER approach. The RER approach generated 73, as opposed to 33 AOCM nodes since only accurate codes were displayed (Table 4).

Table 4
The word co-occurrence network generated via the semi-automated (RER) approach produced a slightly denser network, coding for more concepts, and with more connection among concepts in comparison to the automated (AOCM) generated network.

	AOCM	RER
Node count	33	73
Clustering coefficient		
Network density	0.312	0.338
Node average	0.312	0.338
Standard deviation	0.365	0.180
Min	0.000	0.000
Max	1.000	0.619
Betweenness		
Network centralization	0.403	0.108
Node average	0.085	0.021
Standard deviation	0.117	0.030
Min	0.000	0.000
Max	0.475	0.136
Total degree		
Network centralization	0.034	0.059
Node average	0.012	0.019
Standard deviation	0.011	0.027
Min	0.003	0.001
Max	0.044	0.155

Table 5
Degree centrality for nodes (concepts) coded by the semi-automatic (RER) and automated (AOCM) approaches. A higher degree signifies that the concepts were tied to other coded concepts in the network.

AOCM	Degree	RER	Degree
592 household	0.044	Livestock	0.155
131 location	0.044	Kinship	0.136
632 towns	0.037	Political boundary	0.078
121 theoretical orientation in research and its results	0.027	Biomes and land cover	0.064
721 instigation of war	0.024	Conflict	0.063
613 lineages	0.020	Economy	0.054
443 retail marketing	0.017	Land use	0.043
312 water supply	0.001	Land resource	0.039

The networks were denser in the RER network; that is, there were more connections across concepts, and it was easier for one concept to connect with another concept for the RER approach. This effect can be a simple function of the higher recall rate of this method.

By analyzing the eight most important codes from a network perspective in terms of degree centrality and betweenness centrality, we can see that AOCM and RER picked up on many similar concepts and these concepts had high degree—meaning that they were salient concepts in the analyzed text (Table 5 and Fig. 3). The AOCM network shows *AOCM 529 household* and *AOCM 131 location* as sharing the highest degree measurement. The RER network analysis shows that *RER livestock* had the most ties to other concepts in the text (high degree centrality), while the AOCM had no similar corresponding code in its network. *ACOM 613 lineages* and *AOCM 592 household* may be representative of the *RER kinship*. *AOCM 632 towns* and *AOCM 131 locations* may represent *RER political boundary*. *AOCM 721 instigation of war* is similar to *RER conflict*. *AOCM 443 retail marketing* is similar to *RER economy*. *ACOM 121 theoretical orientation in research and its results* had no similar code in the RER network because this code was not represented in the RER content dictionary. The RER approach picked up on more environmental concepts (*RER Biomes and Land Cover* and *RER Land Resource*), and these

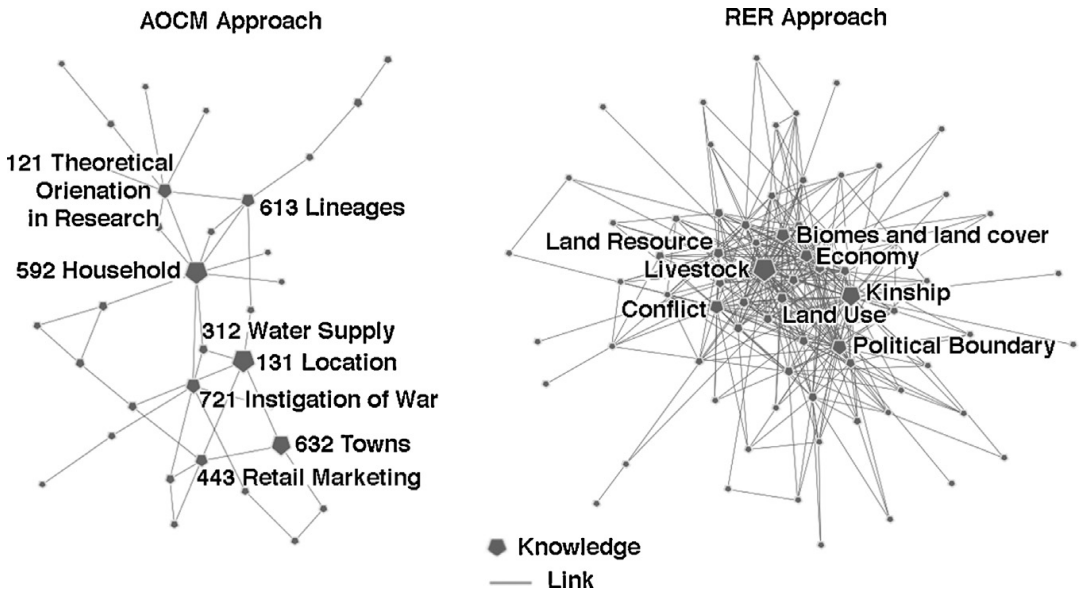

Fig. 3. Main components for the automated (AOCM) and semi-automated (RER) approaches. Concept nodes are sized by degree centrality.

Table 6
Betweenness centrality for nodes (concepts) coded by the semi-automatic (RER) and automated (AOCM) approaches. A higher betweenness centrality signifies that the concept is in a unique position because it bridges together other coded concepts in the network.

AOCM	Betweenness	RER	Betweenness
592 household	0.475	Livestock	0.136
121 theoretical orientation in research and its results	0.356	Kinship	0.114
721 instigation of war	0.332	Land resource	0.105
613 lineages	0.240	Political boundary	0.090
312 water supply	0.229	Economy task	0.083
131 location	0.187	Agriculture	0.073
443 retail marketing	0.158	Intermediate conflict task	0.070
162 composition of population	0.121	Biomes and land cover	0.061

concepts had a high degree centrailty, meaning that they were salient in the text; however, these terms were not picked up on by the AOCM approach.

In both networks, kinship, political boundaries, and intermediate conflict were important bridging concepts (Table 6 and Fig. 4). Of course, since the AOCM did not code for livestock that concept did not appear as an important bridging component, nor did any of the environmental terms, in contrast to the RER analysis. The *AOCM 162 composition of population* and *RER agriculture* ranked high in betweenness centrality metrics (as opposed to degree centrality metrics).

Terms associated with livestock (animal-by-products, bone, horn and shell technology) do not appear highly central in the AOCM approach. The AOCM approach may have more false positives, but most of the major concepts and main points appear in both the AOCM and RER approaches using the degree and betweenness centrality metrics that focus in on the most important topics. On the other hand, livestock, which was an important bridging concept, was not picked up on by the automated approach and this can be remedied by coding for these terms.

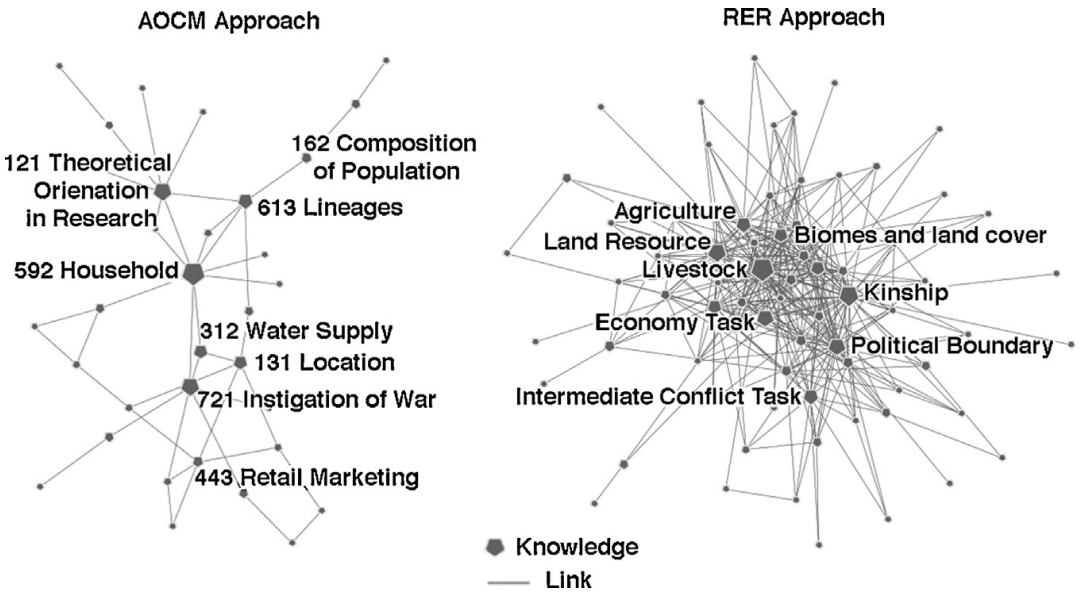

Fig. 4. Main components for the automated (AOCM) and semi-automated (RER) approaches. Concept nodes are sized by betweenness centrality.

4. Discussion

The large amount of time needed to develop the semi-automated RER code scheme is necessary because the RER approach had the highest balance of recall and precision. The RER codes matched similar concepts that the human coder found (88%). In contrast, the automatically generated dictionaries can likely pick up on broad concepts, but need to be refined for more precise coding. This finding agrees with prior research which has shown that, even though network metrics run on matrices extracted with different methods from the same underlying text data, the agreement in highest scoring key players can still be high (Diesner, 2012). This prior work has also shown that the automated mapping of text terms to ontological categories—as provided in AutoMap and used herein—generalizes strongly across different genres, with the genres considered being email data, news wire data and funded research proposals (Diesner, 2012). Technologies that support creation of codes like the RER are needed, and indeed, due to this work were developed and have been shown to reduce coding time significantly (Carley et al., 2011b). The systematicity of the AOCM and RER are such that they generate on average more consistent coding than will human coding, particularly for large corpora. Another feature of this work is that the RER codes are not specific to Somalia or this data and can be re-used. Currently, they are being used to code a large corpus of texts (all news articles for eight years from the *Sudan Tribune* online newspaper; Van Holt et al., 2012). However, adaptation to other genres and content domains might still require additional human labor.

As noted above, we added new coding terms. We have referred to these as RER terms. These were added to the original OCM codes for the simple reason that technology, social evolution and advances in science leading to the discovery of new drivers of cultural change had led to factors of interest that we wanted to code for that were not part of the original set of the OCM codes. We note that this is likely to be the norm and that a canonical set of comprehensive terms is unlikely to emerge in the near future. This makes the need for technologies that support the rapid human-in-the-loop coding, like RER, critical. Moreover, machine learning methods, in general, and statistical natural language processing methods, in particular—such as boot-strapping and semi-supervised domain adaptation—have shown to be useful to adopt existing dictionaries to new domains and data sets with no to minimal human intervention (Daumé, 2007; Gupta and Sarawagi, 2009).

4.1. Error analysis and word sense disambiguation

Because each automated code (AOCM and RER) picked up on a word or group of words, a word can be taken out of context and inappropriately classified. As of now, our dictionary can only code to one concept and word sense disambiguation is a concern (see the example below), and they account for most of the false positives. For example, *RER ecological concepts* was coded for in three instances, each time coding for **nature**, but the text was not in an ecological context.

> The fluid and pragmatic *nature* of Somali politics... (text 33)
> ...the fragmented *nature* of the Somalis' historical experience... (text 38)
> One concerns the *nature* of the Somali "conquest" of the Horn of Africa... (text 46)

Also *RER oil*, appeared once thought the texts but incorrectly coded for **reserves**, which in the document referred to animal reserves and not oil reserves.

> ... access to dry season grazing *reserves... (text 33)*

The *RER property* incorrectly coded for **will**. A will may be considered property if it refers to a last will and testament, but in the majority of cases it will be used as a verb to proclaim future action or intention, such as it does in the analyzed texts.

> In later chapters I *will* describe... (text 35)
> "If a man (of the town) lives long enough, he *will* get to see everything... (text 42)
> Lambs conceived on that night *will* be born about 150 days later... (text 71)

One possibility would be to use linguistics to provide dual meanings. We could ask the computer to search for **nature of** and code that to another concept, such as RER-characteristics (a new category) or to move **nature of** to the delete list for this analysis. Another possibility is to use **a will** to code for the document that is related to property. Also, AutoMap can analyze parts of speech, which helps to disambiguate terms. For example, **will** can be a noun (in the sense of a document) or a verb. For a human coder, this disambiguation needs to be done on a per concept basis because the user must go back to the original text and see how well the analysis reflects the meaning of the text.

Of the 98 correct RER codes, 19 had both correct and incorrect coding instances. The *RER fauna* accounted for the most incorrect coding occurrences from a single RER, with ten incorrect and only eight correct. In every instance of incorrect coding **horn**, coded to *RER fauna*, but in a context referring to the horn of Africa. **Horn of Africa** could be coded to *RER boundary* and **horn** could remain *RER fauna*. If such a change is made, it is important to program the software so **horn of Africa** is searched for first and **horn** is searched for second.

4.2. Network analysis

The network analysis on the resulting data quantifies how the concepts are related. For general concepts, the AOCM and RER network analysis showed similar findings. If this RER coding system and network analysis was used to index HFRAF texts, cross cultural texts could be sampled by how salient a concept was in a text or how a concept was related to another code. So, we could find out not only which texts discussed kinship, but also we could see the relationships between kinship and other concepts. We then could analyze the structure of those texts with the RER to ontology classification to see how the structure of society is distinct and relates to ecological and geospatial attributes, for example. Finally, we could evaluate which texts had kinship as more or less central concepts (degree centrality) and when kinship was a key part of society that bridged together other parts of society (betweenness centrality). In fact, new types of ethnological studies could be developed by creating units of analysis for comparison based on the relationships among and within cultures, societies, ethnic groups, etc., rather than the characteristics of a single culture or subgroup alone.

4.3. Conclusions

Content analysis of newspaper articles, blogs, and other text data resources offers social scientists a new approach for the rapid ethnographic assessment of text-based data sources, where large-scale, over time data can be integrated with other concepts that address human behavior. The considered automatically generated dictionary (AOCM) approach had a high recall rate (many items coded for) but low precision (many false positives) in contrast to the semi-automated RER approach that had few false positives and matched the gold standard human relations area files (OCM). The RER approach, however, had higher recall, picked up on even more concepts,

than the human coders (OCM). The main false positives in the RER analysis were word sense disambiguation issues, which were words that coded for more than one concept. What does that mean for other text coding projects? When large data sets are available, precision is often more crucial than recall, which makes RER a more suitable approach. When only a small data set is to be coded, recall is often considered more important than precision—which means that AOCM might be a useful strategy, but will require manual refining of the generated dictionary. As another remedy, we hope to use parts of speech to code for some of these issues in the future. When comparing only the correctly coded AOCM and RER via a network approach, in general, both coding schemes picked up on similar topics in a broad sense (top eight topics as per betweenness and degree centrality network measures). Network metrics help to characterize salient concepts across texts. Although we unioned separate ontologies for purposes of comparison, our approach allows for us to visualize and analyze multiple concepts—such as knowledge, resources, and tasks—and therefore integrate data from multiple disciplines over space and time. The network metrics and visualization can transform how ethnological studies are conducted, by basing the units of analysis for comparison on relationships among and between cultures in addition to characteristics within cultures analyzed. The Internet resources and other digitized textual data sources allow us to view data over time at a finer resolution. In combination, this offers new analytical pathway for improving our understanding of human behavior with the use textual sources of data.

Acknowledgements

This work is supported, in part, by the Office of Naval Research (ONR), United States Navy (ONR MURI N000140811186). Additional support was provided by the Center for Computational Analysis of Social and Organizational Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research, or the U.S. government.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.poetic.2013.05.004.

References

Abello, J., Broadwell, P., Tangherlini, T.R., 2012. Computational folkloristics. Communications of the ACM 55 (7) 60–70.

Bernard, H.R., Ryan, G.W., 2009. Analyzing Qualitative Data: Systematic Approaches. Sage, Thousand Oaks, CA.

Boster, J.S., Johnson, J.C., 1989. Form or function: a comparison of expert and novice judgment of similarity among fish. American Anthropologist 91, 866–889.

Bourdieu, P., Wacquant, L.J.D., 1992. An Invitation to Reflexive Sociology. University of Chicago Press, Chicago.

Brandt, P.T., Freeman, J.R., Schrodt, P.T., 2011. Real time, time series forecasting of inter- and intra-state political conflict. Conflict Management and Peace Science 28 (1) 40–63.

Carley, K.M., 1994. Extracting culture through textual analysis. Poetics 22, 291–312.

Carley, K.M., Diesner, J., Reminga, J., Tsvetovat, M., 2007. Toward an interoperable dynamic network analysis toolkit: decision support systems (Special Issue on Cyberinfrastructure for Homeland Security). Advances in Information Sharing, Data Mining, and Collaboration Systems 43 (4) 1324–1347.

Carley, K.M., Columbus, D., Bigrigg, M., Kunkel, F., 2011a. AutoMap User's Guide 2011. Technical Report, CMU-ISR-11-108. Carnegie Mellon University, School of Computer Science, Institute for Software Research.

Carley, K.M., Reminga, J., Storrick, J., Columbus, D., 2011b. ORA User's Guide 2011. Technical Report, CMU-ISR-11-107. Carnegie Mellon University, School of Computer Science, Institute for Software Research.

Carley, K.M., Bigrigg, M.W., Diallo, B., 2012. Data-to-model: a mixed initiative approach for rapid ethnographic assessment. Computational and Mathematical Organization Theory 18 (3) 300–327.

Cassanelli, L.V., 1982. The Shaping of Somali Society: Reconstructing the History of a Pastoral People, 1600–1900. University of Pennsylvania Press, Philadelphia.

CRA, 2001. Analyses of News Stories on the Terrorist Attack. Available at: http://locks//locsks.asu.edu/terror.

Danowski, J.A., 1993. Network analysis of message content. Progress in Communication Sciences 12, 198–221.

Daumé, H., 2007. Frustratingly easy domain adaptation. In: Proceedings of 45th Annual Meeting of the Association of Computational Linguistics (ACL), Prague, Czech Republic, pp. 256–263.

De Nooy, W., 2003. Fields and networks: correspondence analysis and social network analysis in the framework of field theory. Poetics 31, 305–327.

Diesner, J., 2012. Uncovering and Managing the Impact of Methodological Choices for the Computational Construction of Socio-Technical Networks from Texts. Technical Report, Carnegie Mellon CMU-ISR-12-101. (Ph.D. Thesis).

Diesner, J., Carley, K.M., 2008. Conditional random fields for entity extraction and ontological text coding. Journal of Computational and Mathematical Organization Theory 14, 248–262.

Diesner, J., Carley, K., 2010. Extraktion relationaler daten aus texten.(Relation extraction from text). In: Stegbauer, C., Häußling, R. (Eds.), Handbuch Netzwerkforschung. (Handbook of Network Research).Vs Verlag, Weisbaden, pp. 507–521.

Gerner, D., Schrodt, P., Francisco, R., Weddle, J., 1994. Machine coding of event data using regional and international sources. International Studies Quarterly 38 (1) 91–119.

Godoy, R., Reyes-Garcıa, V., Byron, E., Leonard, W., Vadez, V., 2005. The effect of market economies on the well-being of indigenous peoples and on their use of renewable natural resources. Annual Review of Anthropology 34, 121–138.

Gupta, R., Sarawagi, S., 2009. Domain adaptation of information extraction models. ACM SIGMOD Record 37 (4) 35–40.

Holsti, O.R., 1969. Content Analysis for the Social Sciences and Humanities. Addison-Wesley, Reading, MA.

Johnson, J.C., Griffith, D.C., 1998. Visual data: collection, analysis, and representation. In: de Munck, V., Sabo, E. (Eds.), Using Methods in the Field: A Practical Introduction and Casebook. Altamira Press, Walnut Creek, CA, pp. 211–228.

Johnson, J.C., Krempel, L., 2004. Network visualization: ''The Bush Team'' in Reuters News Ticker 9/11-11/15. Journal of Social Structure 5 (4) .

Johnson, J.C., Orbach, M.K., 2002. Perceiving the political landscape: ego biases in cognitive political networks. Social Networks 24, 291–310.

King, G., Lowe, W., 2003. An automated information extraction tool for international conflict data with performance as good as human coders: a rare events evaluation design. International Organization 57 (3) 617–642.

Krinsky, J., 2010. Dynamics of hegemony: mapping mechanisms of cultural and political power in the debates over workfare in New York City, 1993–1999. Poetics 38, 625–648.

Mack, A., 2007. Global Political Violence: Explaining the Post-Cold War Decline, Coping with Crisis. Working Paper Series. International Peace Academy, New York.

Murdock, G.P., 1983. Outline of World Cultures, 6th edition. Human Relations Area Files, New Haven, CT.

Murdock, G.P., 1987. Outline of Cultural Materials, 5th edition. Human Relations Area Files, New Haven, CT.

Murdock, G.P., White, D.R., 1969. Standard cross-cultural sample. Ethnology 8, 329–369.

National Research Council, 2008. Behavioral modeling and simulation: from individuals to societies. Committee on Organizational Modeling from Individual to Societies (G.L. Zacharias, J. McMillan, H. Arrow, S.P. Borgatti, R. Burton, K.M. Carley, C. Dibble, E. Hudlicka, J.C. Johnson, S.E. Page, A. Sage, L.S. Tesfatsion, and M.J. Zyda). In: Zacharias, G.L., McMillan, J., Van Hemel, S. (Eds.), Board on Behavioral, Cognitive, and Sensory Sciences, Division of Behavioral and Social Sciences and Education. The National Academies Press, Washington, DC.

Neuendorf, K.A., 2002. The Content Analysis Guidebook. Sage, Thousand Oaks, CA.

O'Reilly, T., 2005. What is Web 2.0? Design Patterns and Business Models for the Next Generation Software. Available at: http://oreilly.com/web2/archive/what-is-web-20.html.

Osgood, C., 1959. Suggestions for winning the real war with communism. Journal of Conflict Resolution 3, 295–325.

Reyes-Garcia, V., Byron, E., Vadez, V., Godoy, R., Limache, E.P., Leonard, W.R., Wilkie, D., 2004. Measuring culture as shared knowledge: do data collection formats matter? Cultural knowledge of plant uses among Tsimane' Amerindians, Bolivia. Field Methods 16 (2) 135–156.

Rinner, D., Kebler, C., Andrulis, S., 2008. The use of Web 2.0 concepts to support deliberation in spatial decision-making. Computers, Environment, and Urban Systems 3, 386–395.
Roberts, C.W., 1997. A generic semantic grammar for quantitative text analysis: applications to East and West Berlin radio news content from 1979. Sociological Methodology 27, 89–129.
Schneider, G., Geditsch, N.P., Carey, S., 2011. Forecasting in international relations: one quest, three approaches. Conflict Management and Peace Science 28 (1) 5–14.
Schrodt, P.A., Gerner, D.J., 1994. Validity assessment of a machine-coded event data set for the Middle East, 1982–92. American Journal of Political Science 38 (3) 825–854.
Spradley, J., 1979. The Ethnographic Interview. Wadsworth Group, Belmont, CA.
Stone, P.J., Bales, R.F., Namenwirth, J.Z., Ogilvie, D.M., 1962. The General Inquirer: a computer system for content analysis and retrieval based on the sentence as a unit of information. Behavioral Science 7, 484–498.
van Cuilenburg, J., Kleinnijenhuis, J., de Ridder, J., 1986. A theory of evaluative discourse: towards a graph theory of journalistic texts. European Journal of Communication 1 (1) 65–96.
Van Holt, T., 2012. Landscape influences on fisher success: adaptation strategies in closed and open Access fisheries in Southern Chile. Ecology and Society 17 (1) 28–44.
Van Holt, T., Johnson, J., Brinkley, J., Carley, K., Caspersen, J., 2012. Structure of ethnic violence in Sudan: a semi-automated network analysis of online news (2003–2010). Computational and Mathematical Organization Theory 18 (3) 340–355.
Whiting, J.W.M., 1986. George Peter Murdock (1897–1985). American Anthropologist 88 (3) 682–686.

Dr. Tracy Van Holt's interests include human–environment interactions as they relate to natural resource use, the consequences of landscape and environmental change, climate change, conflict, and sustainable development. She works with communities and research questions in tropical and temperate ecosystems as well as wetland and coastal environments. She integrates geospatially explicit remotely sensed data with social and environmental data.

Dr. Jeffrey C. Johnson's interests include the influence of technological and environmental factors on the organization of work, leisure, and cognition, particularly in groups in extreme and isolated environments. He has focused a major portion of his teaching and research program around the use of social network theories and methods for understanding social structure and organization. His recent substantive interests have focused on the relationship between cognition and social structure. The bulk of his research has focused on these concerns among the maritime peoples of the Pacific basin, especially the insular Central Pacific, the Caribbean, and coastal North America. Interdisciplinary in both training and orientation, he has had teaching experience in economics, anthropology, sociology, statistics, and Pacific studies.

Dr. Kathleen M. Carley's interests include dynamic network analysis, computational, social and organization theory, adaptation and evolution, text mining, and the impact of telecommunication technologies and policy on communication, information diffusion, belief evolution, disease contagion and response within and among groups particularly in disaster or crisis situations. In her research, she combines results and approaches from cognitive science, organization science, social networks and computer science to address complex social and organizational problems. She and her Center for Computational Analysis of Social and Organizational Systems (CASOS) have developed advanced technologies for network analytics and visualization (ORA), Network extraction from texts (AutoMap), and network evolution and diffusion (Construct and BioWar).

James Brinkley is a PhD student in the Coastal Resource Management program and a graduate research assistant for the Institute for Coastal Science and Policy at East Carolina University. His interests include coastal hazards, cultural knowledge of maritime communities, coastal and environmental planning, and resource conflict issues. Current work in his interdisciplinary academic program involves using various social science and geospatial analyses to study cultural understanding of coastal natural hazards.

Jana Diesner is an Assistant Professor at the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign. Jana conducts research at the nexus of network analysis, natural language processing and machine learning. Her goal is to contribute to the computational analysis and better understanding of the interplay and co-evolution of information and networks. She develops and analyzes methods and technologies for extracting network data from text corpora and considering the content of information for network analysis. She studies networks from the business, science and geopolitical domain, and is particularly interested in covert information and covert networks.