Comput Math Organ Theory (2012) 18:300–327
DOI 10.1007/s10588-012-9125-y

SI: DATA TO MODEL

Data-to-model: a mixed initiative approach for rapid ethnographic assessment

Kathleen M. Carley · Michael W. Bigrigg · Boubacar Diallo

Published online: 18 July 2012

Abstract Rapid ethnographic assessment is used when there is a need to quickly create a socio-cultural profile of a group or region. While there are many forms such an assessment can take, we view it as providing insight into who are the key actors, what are the key issues, sentiments, resources, activities and locations, how have these changed in recent times, and what roles do the various actors play. We propose a mixed initiative rapid ethnographic approach that supports socio-cultural assessment through a network analysis lens. We refer to this as the data-to-model (D2M) process. In D2M, semi-automated computer-based text-mining and machine learning techniques are used to extract networks linking people, groups, issues, sentiments, resources, activities and locations from vast quantities of texts. Human-in-the-loop procedures are then used to tune and correct the extracted data and refine the computational extraction. Computational post-processing is then used to refine the extracted data and augment it with other information, such as the latitude and longitude of particular cities. This methodology is described and key challenges illustrated using three distinct data sets. We find that the data-to-model approach provides a reusable, scalable, rapid approach for generating a rapid ethnographic assessment in which human effort and coding errors are reduced, and the resulting coding can be replicated.

Keywords Text-mining · Network-analysis · Meta-network · Social-networks · Agent-based simulation · Data analysis · Newspaper data

K.M. Carley (✉)
Wean 5130, ISR, SCS, Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: kathleen.carley@cs.cmu.edu

M.W. Bigrigg · B. Diallo
Carnegie Mellon University, Pittsburgh, PA 15213, USA

M.W. Bigrigg
e-mail: bigrigg@cs.cmu.edu

B. Diallo
e-mail: b.diallo81@gmail.com

1 Introduction

Whether the concern is providing humanitarian assistance, responding to a disaster, curtailing civil violence, exploring the impact of a new law, drug, or social policy, or assessing the political landscape, researchers, analysts and policy makers often find themselves needing to rapidly understand "the lay of the land" and the recent changes in the socio-cultural environment. Central questions of importance center around who do you talk with to get things done, and how can the relevant groups be influenced. The ability to rapidly address these and other socio-cultural questions is crucial. Moreover, in many environments, there is a need to address these questions remotely; i.e., without actually going to the region of concern or meeting with the members of the group.

Rapid ethnographic assessment is an approach for rapidly characterizing the socio-cultural landscape (Beebe 1995). Rapid ethnographic assessment provides insight into the socio-cultural nature of a region by identifying the current key actors, issues, sentiments, resources, activities and locations, and any recent changes. Rapid ethnographic assessment is particularly useful when product demands allow little time for detailed anthropological field work (Bauersfeld and Halgren 1996). It is also called for when groups are faced with operating in a region where they have little experience or working with an unfamiliar group (Bentley et al. 1988).

Traditional ethnography requires massive amounts of painstaking detailed daily investigation into a region or group and can take years. The result is a detailed rich understanding that can support theory development and model testing. The results are valuable and necessary for a thorough and detailed understanding of a region or group; or for creating a history of that socio-cultural milieu. However the approach does not scale and the specific methods used are often context specific. In contrast, rapid ethnographic assessment though it derives from anthropological fieldwork favors reproducibility, speed, and general understanding over exquisite detail. Rapid ethnographic assessment is increasingly widely used in a number of fields from human computer interaction to medicine to counter-terrorism. However, the set of techniques vary widely. Common themes are key informants, rapid assessment, and field data. We suggest that this process can be enhanced through the use of computer assisted remote assessment employing documents about or produced in or by the region or group of interest. Moreover, we suggest that such a technology would not only be a key tool for ethnographers, but will admit overall improvements in replicability and enable scalable assessment technology to be developed.

Herein, we propose, and demonstrate, a computer-assisted data-to-model approach for rapid ethnographic assessment. The proposed approach is a complement to, not a replacement for, traditional ethnography; and serves to fill a key gap—rapid remote assessment. The approach we present is centered on the idea that texts contain information that can be extracted and codified as a network linking actors, issues, sentiments, resources, activities and locations. By examining changes in these networks, the current and evolving socio-cultural landscape can be characterized. While the assessment is not as rich as a detailed ethnography, it does provide guidance as to the main points of interest be they key actors or contentious issues.

The proposed Rapid Ethnographic system as embodied in the data-to-model (D2M) process supports analysts in assessing change in the socio-cultural system

(Carley et al. 2011a, 2011b). The D2M process can make use of open-source text data, gazateers, twitter feeds, scholarly articles, and other data available electronically. From this over-arching corpus meta-network (Carley 2002; Krackhardt and Carley 1998) information is extracted; i.e., who, what, where, when, why are extracted as are the relations among them. Geo-temporal time-stamping and attributes of nodes and links provide additional empirical value. Extracted data is processed, basic network analyses are run, and data is output to simulations for simple forecasts.

The heart of this approach is open-source text exploitation. The tremendous growth in the amount of digital information available, largely in unstructured or text form makes this approach possible. In creating ethnographies, most researchers assess open-source text data by either reading samples of the information or through extremely labor intensive approaches to coding texts. The sheer volume of texts suggests the need for a new, more automated, approach (Alexa 1997). In contrast to a fully automated approach, we use a mixed-initiative approach employing machine learning and text mining techniques and human-in-the-loop refinement for rapidly processing vast quantities of text-based information to provide the analyst with a high level understanding of the socio-cultural system in the region of interest and the way it has recently evolved. Semi-automated approaches are efficient when dealing with vast quantities of unstructured texts which cannot be analyzed individually by Subject Matter Experts (SME) in a timely, unbiased and systematic fashion (Corman et al. 2002).

The overall text analysis approach is often referred to as Network Text Analysis (NTA) which encodes a meta-network based on the relationships of words in a text (Carley 1997; Popping 2000) and then comparing the network of texts (Batagelj et al. 2002). We note that within the field of text-mining there exist a number of techniques each operating independently and researched separately (Manning et al. 2008). These include, but are not limited to entity extraction (Chakrabarti 2002), content analysis (Holsti 1969; Krippendorff 2004), topic modeling (Hofmann 1999; Blei et al. 2004), entity resolution and event discovery (Roth and Yih 2007), theme analysis (Landauer et al. 1998), role discovery (Wang et al. 2010), link extraction (Ramakrishnan et al. 2006), multi-dimensional extraction (Lin et al. 2010; Zhang et al. 2009; Ding et al. 2010) and meta-network extraction (Carley et al. 2011b; Diesner and Carley 2008). Our approach employs a number of these individual techniques, and other techniques not heretofore mentioned such as deduplication and anaphor resolution. The suite of techniques are placed into an integrated and expanded workflow in which we have optimized the ordering of applications so as to minimize human effort and time to develop extracted networks. Overall, the data-to-model process employs a vast number of distinct methodologies from web-scrapers and sensor feeds to text-mining techniques to network analysis tools to agent-based simulations. Moreover, it is extensible whereby new techniques can be added as they are developed.

Central to the data-to-model approach is network analysis. Specifically, this approach involves the extraction, analysis, and then simulation of the impact of various interventions of a network based model of the complex socio-cultural system being analyzed. The system is represented as a meta-network (Carley 2002, 2006) in which an ecology of networks linking the who, what, when, where, why and how are collec-

tively incorporated into a multi-mode, multi-link, multi-level network representation system.

In this paper, we describe the data-to model process that enables meta-network information to be extracted from vast quantities of unstructured texts-based information in an efficient manner while minimizing the strain on the human coders and increasing the overall effectiveness of the extraction and coding process. We describe the methodology for the development process. In addition, we describe the central text mining and surrounding data cleaning methodology. Other methodologies, such as web-scrapers and agent-based models are not described. We have identified what the preferred coding choices are so as to create a more automated system, and have put the entire process into a more detailed workflow so as to reduce analyst time. These issues, our solution, and key features of the workflow are described. We focus here on the data encoding step and the way in which SME input plays an integral role in that process. We demonstrate the overall workflow that moves from text collection to data extraction to analysis. The impact of the process on data is demonstrated using three data corpi. Information on time savings and nature of data extracted is provided.

Human involvement often that of subject matter experts, is critical to the overall process. From a development perspective, humans played the following roles: (a) developed specialized generalization thesauri, (b) vetted computer generated thesauri, (c) vetted extract networks for people to people, ethnic-group/tribe to ethnic-group/tribe, people to ethnic-group/tribe, organization to organization, people to organization, (d) vetted time line, and (e) confirmed results. We found that, given the current state of text mining and gazetteers, there is a need for humans in the initial thesauri construction effort to disambiguate some place names from people/organizational names. The confirmation of results was done by (a) having humans review the results and identify surprises, inaccuracies, or known missing information and (b) presenting the results to regional experts and collecting feedback on surprises, inaccuracies, or known missing information that was then incorporated into the coding. We also tracked the time that humans spent engaged in the coding activities with and without this process, thereby providing guidance on the value added of the data-to-model process. This information is presented herein.

2 Data

We tested the data-to-model process using three corpi:

Sudan: Social change and situation assessment. 71,000 texts from news-sources, websites, books, and writings by scholars. Project-based thesauri contains 38,552 gazetteer locations and 16,001 entries.

Afghanistan: primary and secondary actors and region of influence analysis: 247 large texts from news-sources and websites. Project-based thesauri contains 146,573 entries.

Catnet: competitive adaptation analysis: 1,109 texts from news-sources and websites. Project-based thesauri contains 28,727 entries.

3 Data-to-model (D2M) process

The D2M process mirrors a manual approach that is taken by the analyst for developing a model from unstructured text data: downloading and fixing data, cleaning and preparing text, finding important concepts, forming relationships between concepts, attempting initial analysis, further refinement of concepts and relationships with associated re-analysis, and then finally after the output is correct the addition of attributes to the concepts, and final analysis results. Using D2M the analyst converts raw texts to networks; specifically to meta-networks. In addition a semantic network and a project thesauri are developed.

A semantic network is a set of interlinked concepts. Using D2M, concepts are extracted and then linked into a semantic network based on semantics and syntactic proximity, e.g., co-occurrence within a window of two sentences in length. Additional inferencing based on part-of-speech and type of concept is further used to refine the connection.

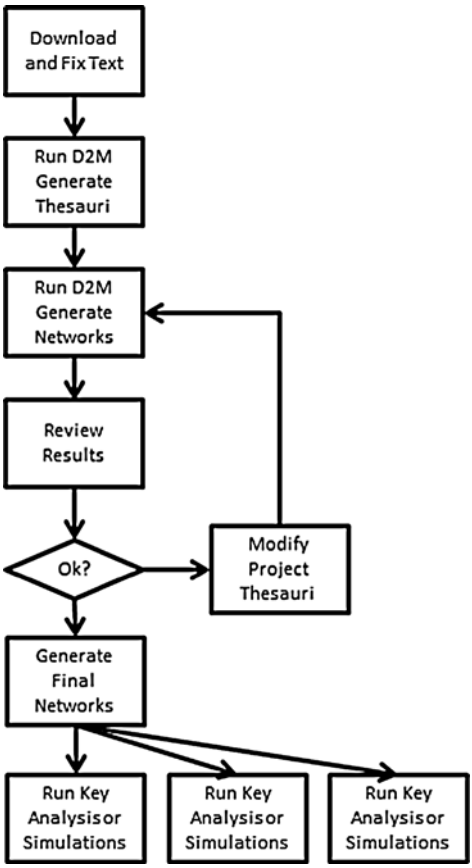
A meta-network is a geo-temporal set of interlinked networks connecting nodes in multiple ontological classes. The ontological classes we use are: agents, organizations, locations, knowledge, resources, beliefs, tasks and events. Using D2M, concepts are extracted, linked into a semantic network, cross-classified into their ontological class, and attributes of nodes are added. Further, using D2M alternate "versions" of the same basic concept are collapsed into the same term.

Two types of thesauri are used by D2M. The first is a generalization thesauri in which all concepts referring to the same ideological kernel are converted to a common term; e.g., all aliases for an individual are converted into a common first-name_lastname format. The second type of thesauri is an ontology thesauri in which the concepts are cross-classified into their ontological type. The types currently supported are agent, organization, location, task, knowledge, resource, belief and event. In addition, this latter type of thesauri supports a second level of classification into specific and generic for agents, organizations and locations.

The D2M process is designed to minimize the strain on the analyst of using network text analysis to generate this meta-network model and the difficulty in using a multi-tool cycle to clean the meta-network model. D2M uses various component tools and routines in AutoMap (Carley et al. 2011b) and ORA (Carley et al. 2011a) in predefined sequences. The key to D2M is that the component tools are used in a prescribed manner consistent with the workflow that would be executed by a well trained expert analyst to process unstructured textual data into a network model. D2M is a mixed initiative process; i.e., it employs automated processing which is augmented and refined through a human-in-the-loop cleaning process.

In order to use text-mining techniques to support modeling the analyst has to go through a large number of steps. Figure 1 shows a typical workflow that outlines, at a high level, the basic steps that an analyst will follow to develop a model from a data corpus of unstructured text. We have developed a data-to-model procedure that moves directly from identification of relevant texts to model-based forecasting, minimizing strain on the analyst and supporting overall systematic data coding. Based on numerous assessments of human analysts engaged in extracting and processing network information from texts, we identified a common workflow and operationalized

Fig. 1 Analyst workflow



it into a simple system which directs the analyst through a process of data processing and thesauri construction.

3.1 Component technologies

At the core of the data-to-model process are two basic sub-tools, AutoMap and ORA. AutoMap and ORA were designed to work together, i.e., the output of AutoMap is directly readable by ORA. Nevertheless, for most analysts, determining how to use AutoMap to generate data for ORA and then cleaning that data by going back and forth between AutoMap and ORA can be daunting as it involves making a large number of coding choices. Through interviewing analysts engaged in this process, observing students learn this process, and testing this process ourselves we identified a common workflow. In addition, review of the literature and analyst observation led to the identification of a set of best practices in terms of parameter settings. This information was codified into the D2M workflow. D2M can be supported through SORASCS (Garlan et al. 2009), or locally through a script-runner which is part of AutoMap (Carley et al. 2011b). SORASCS is a workflow management system for

services and thick clients. AutoMap and ORA as thick clients, and variations of their components as services are in SORASCS.

AutoMap (Carley et al. 2011b) is used for text-mining and data extraction. AutoMap is a toolbox of techniques and approaches for network text analysis. It contains 24 routines to preprocess (remove concepts, find-and-replace concepts and phrases, stemming), 28 routines to generate results (extract named entities, nouns/verbs, create meta-networks and semantic networks), 44 routines to do post-processing (adding attributes), and 18 other supplemental routines (removing headers, removing HTML symbols, extract text from documents). In total there are 114 routines available for the analyst. The selection of routines and the order of operation of these routines require a person familiar with network text analysis to understand the implications of the choice and order of the routines. AutoMap can operate as a thick client system or its component tools can act as services inside of and be called from a service oriented architecture like SORASCS. In order to go from raw text to the extracted data, which can be thought of as a model of the situation, the analyst must determine which of these routines to use, and in what order. A daunting task!

ORA (Carley et al. 2011a) is used for network analytics and simulation. ORA is a toolbox of techniques for network analysis. It contains 3 analysis engines (including LSA, latent semantic analysis), 44 reports (key entity identification and geospatial assessment), 157 measures (all SNA, social network analysis, measures and additional custom measures), 3 simulation systems (such as near term analysis), and 16 visualization systems (including viewing measures and networks over time). ORA can operate as a thick client system or its reports and visualizers can act as services inside of and can be called from a service oriented architecture like SORASCS.

Given the extracted model from AutoMap the analyst now needs to examine it with ORA; but in doing so may find errors. A typical error is that there are multiple spellings of the same person's name. The analyst now needs to engage in a process of data cleaning, which can involve going back and forth between AutoMap and ORA anywhere from two to ten times.

3.2 High level overview of D2M steps

At a high level, D2M involves seven basic steps:

1. Corpus Collection
2. Corpus Cleaning
3. Meta-Network Extraction
4. Refinement (Thesauri Construction and Cleaning)
5. Final Meta-Network Extraction
6. Data Augmentation
7. Network Analytics and Forecasting

These activities are linked to logical steps in Fig. 2. Step seven can be quite broad and includes both descriptive and predictive network analysis, geo-network analysis, intervention identification and forecasting and assessment using agent-based dynamic-network simulation tools.

Many of the semi-automated processing steps are invisible, though not hidden from the analyst. The automated D2M process requests a minimum of information from the analyst:

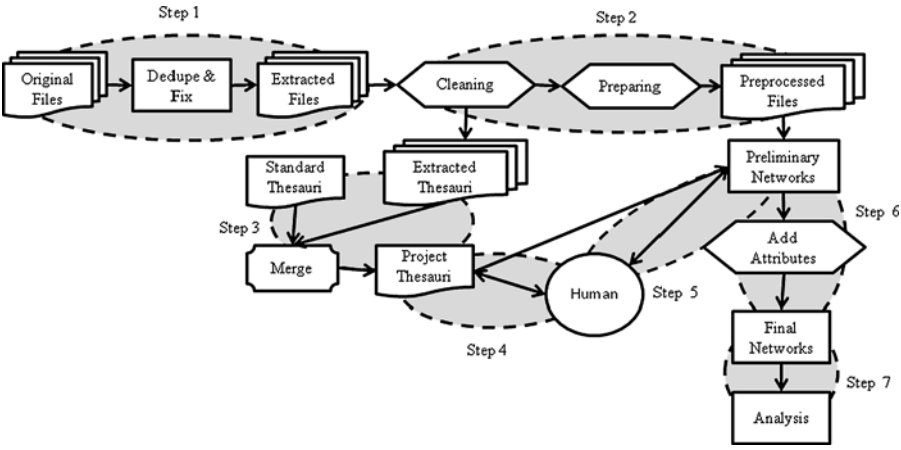


Fig. 2 D2M detailed workflow

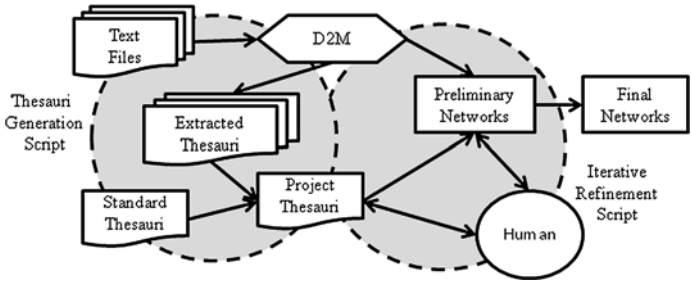


Fig. 3 Workflow scripts

- The location of the text files to be processed
- The location of a directory that can be used for intermediate results to be reviewed
- Pre-established project-agnostic thesauri
- Project-based thesauri
- Additional attributes thesauri

The analyst can go through the seven steps in the process using two workflow scripts. The first supports the automated generation of the initial meta-networks and an extracted thesauri. The second script is used repeatedly as part of the iterative human-in-the-loop approach to refinement (data cleaning and thesauri augmentation). Figure 3 maps the D2M activities to the two workflow scripts.

The generation of thesauri, script 1, is run once, with the refinement script run as the human-in-the-loop modifies and adapts the project-based thesauri until satisfied. Only three iterations have been needed for most data sets.

Script 1—Meta-network extraction component For data coding, we use machine learning techniques in support of advanced network text analysis. At an abstract and very high level, a network text analysis involves four-steps for converting unstructured texts to meta-networks (Diesner and Carley 2005):

1. Concept Identification: The analyst identifies what concepts are or are not of interest.
2. Entity Identification: The analyst defines an ontology; i.e., the set of entities into which the data will be coded. Typical entity classes are people, organizations, places, resources, tasks, events, and beliefs. Note that step 1 and 2 are temporally interchangeable.
3. Concept Classification: The analyst maps the identified concepts of interest into the entity classes using a meta-network or ontology thesaurus.
4. Meta-network extraction: The analyst using an automated tool, applies the aforementioned thesauri and delete lists to the texts. This tool then processes the texts and extracts concepts, cross-classifies them into their ontological category, and then extracts the relation among the identified concepts. The result is a structured dataset containing a meta-network. Typically there is one such network per text and then these are fused in to a single meta-network.

On the surface this sounds straight forward. It turns out, however, that there are many coding choices that the analyst must make along the way that influence the results (Carley 1993). Moreover, there tend to be multiple iterations, particularly for large datasets. Finally, there are a number of cases where analysts need to be involved in the initial extraction and of theories employed. In creating the D2M, we drew on the elements of this process, automated many of the sub-procedures, and inverted the process so that automated extraction occurred first, and human-in-the-loop processing second. Then we embedded this processing in an overall system with two phases, extraction and cleaning, that can be repeated until the analyst is content with the result.

Script 2—Meta-network regeneration Script 2 is used as part of the cleaning process. Like script 1 it generates a meta-network, semantic network, and thesauri. However, it takes as input the basic project thesauri that is being constructed, and the changes provided by the analyst, for which it updates the thesauri to be used on the next iteration. Then it reapplies this processing to the texts. Note, the cleaning done in script 1 is not redone. That occurs only once in the first script. Instead, this script, script 2, focuses on resolving contradictions in the thesauri construction imposed by analyst changes and reapplies this thesauri.

3.3 Detailed description of D2M steps

3.3.1 Step 1: Corpus collection

The analyst begins with a corpus of raw texts. These texts can be drawn from any number of sources. For the three data sets we used, data was collected via web-scrapers and/or provided by external sources. Meta-information was removed. Duplicate texts were removed.

Data can be provided by a number of sources including emails, news articles, web pages, blogs, twitter or RSS feeds. Many of the data sources include additional unrelated material along with the text. The process of extracting and preparing the data depends highly on the source. A web scraper program can be used to extract

Table 1 Illustration of impact of typical deduplication applied to Sudan data

Sudan	Before	After
Text Files	32,613	18,309
Concepts	88,260	83,150
Average Frequency per Concept	197.42	88.16

web data from web sites. However, web scrapers have to be adjusted to extract the relevant information without also extracting more than is needed or not enough. In spite of having a structured layout online, the actual file may contain header and footer information as well as navigation content, and advertising. Some websites are structured so as to preclude wholesale download of content. News articles provided by a collection service such as Lexis-Nexis may be clean with respect to formatting, but may also contain some semi-structured information such as copyright date and publisher which will be found as part of the text file. Some sources will provide the unstructured text and the semi-structured meta-information separately, minimizing the need to remove the semi-structured information.

The best practice for these texts is to put one text per file. File name should include the date, source, and any identifier that analyst wants to use later for binning the texts into categories. Images should be removed. Some analysts find it useful to remove common meta-data and formatting information.

Depending on the source of the data, deduplication is necessary. The process of deduplication is to remove articles that have the same contents as another file. For example, a collection of email messages may be identical duplicates if the data is collected from many people, all whom were the recipients of the same email message. While it may not seem necessary to deduplicate the files retrieved from web sites or external news agencies, many news articles are close matches in the case of articles that come from a press release. The news is repeated, and put into a style that matches its readership. An in-exact deduplication process will remove additional articles if they are within a threshold similarity for another article.

In the case of the SUDAN corpus, for example, the data was drawn from Lexis-Nexis and the Sudan Tribune review. Hence a number of articles, particularly those by AP wire service were duplicated. Table 1 shows the number of text files before and after deduplication, and identifies the average frequency per concept. Deduplication has a number of advantages. It reduces corpus size thereby speeding up processing. In addition, it reduces bias in the resulting data. It is important to note that deduplication alters the frequency of both certain concepts and certain concept pairs. Changes in the frequency of concepts are shown in Table 2. As such it can and does impact the conclusions that are reached when a network analysis is done on the resulting extracted model at both the semantic network and the meta-network level.

The text files in the corpus that are to be processed are assumed to have been not just collected a-priori to the D2M process but also cleaned of extraction-specific issues. Texts from various sources, e.g. news articles, blogs, scholarly articles, and social media data, all require different tools for pulling them from the web and different human-in-the-loop processing to fix extraction-specific issues such as image formatting, adds, and html tags. AutoMap and SORASCS provide routines to aid

Table 2 Top concepts before and after deduplication and general text reformatting—applied to Sudan data

Before			After		
Concept	Count	%	Concept	Count	%
valencia	1141455	6.55	nanuque	128828	4.60
conflict	867688	4.98	culture	61675	2.20
nanuque	500411	2.87	money	61236	2.19
amp	448679	2.88	badou	38352	1.37
sudan	385560	2.21	conflict	29832	1.06
darfur	244036	1.97	afghanistan	26339	0.94
valence	178629	1.03	mild_conflict	26146	0.93
population	178059	1.02	commander	25917	0.93
faouzi ben mohammed	172782	0.99	or	22006	0.79
badou	152547	0.88	washington	20703	0.74

in this effort, but do not incorporate them into the D2M process, as the order and selection of operations is highly specific to the source of the data.

Both deduplication and the removal of extraction related information do impact the results. For example, as seen in Table 2, there is "amp" as a top concept found in the original text set, which comes from the web internally formatting tag "&" that should be removed from the text. The term amp in other contexts as an abbreviation for the electrical term "ampere" would not be removed. For email and semi-structured data there is often a need to process both the meta-data (e.g., the To, From, Subject fields) with one tool, and the content with another, and then to pairwise combine the results.

3.3.2 Step 2: Corpus cleaning

Corpus cleaning is designed to take the body of the text that the analyst wants to process and augment or alter it to remove various typesetting and linguistic style vagaries. It includes generic preprocessing activities: e.g., typo correction and the expansion of contractions and abbreviations. In addition, during this phase pronouns are resolved and unidentified pronouns removed. This is a completely automated process.

The initial cleaning is actually divided into two phases: syntax-preserving cleaning and syntax-destroying cleaning. In the D2M process, there are steps that require that syntax be preserved as these steps make use of a word's part-of-speech information. One example is the extraction of potential names. Names are found by extracting words that are identified as a proper noun. A proper noun is just one part of speech item that is needed, in addition to verbs and common (non-proper) nouns. The cleaning operations are ordered such that all syntax-preserving operations are done first, the resulting text saved, and then the syntax-destroying operations continue. The final resulting fully cleaned text is used for the network generation and the intermediate clean text is used for the generation of lists used in the construction of a thesauri.

The syntax-preserving operations include: removal of extra spaces, expansion of contractions and abbreviations, typo correction, pronoun resolution, and British to

Table 3 Number of concepts before, then after cleaning, and after preparation

Data Set	Original	After Cleaning	After Preparing
Sudan	110,706	110,772	112,643
Afghanistan	129,875	127,778	130,989
Catnet	29,697	30,224	30,784

American spelling conversion. The removal of extra spaces is not necessary, but similar to the deduplication activity described earlier, it provides benefit during the human-in-the-loop part of the D2M process. The British to American spelling conversion is to provide consistency in spelling between articles which are authored by different individuals. The expansion of contractions and abbreviations will aid in the normalization of the concepts found in the text. Normalization will reduce the number of terms by merging them into similar terms. A common normalization operation is the conversion of Rd into Road. Companies will normalize an address to minimize the number not identical addresses to reduce the cost of mailing catalogs or advertisements, merging "123 Maple Rd" and "123 Maple Road" into a single address. The expansion of contractions and abbreviations will reduce the number of terms that are to be addressed, but increasing a frequency count of the resulting term. The expansion is done with common abbreviations and contractions, such as Rd or USA. The most common project-based abbreviations and contractions are acronyms, which need to be normalized by the analyst. The D2M process will extract potential acronyms for the user to vet, but does not incorporate them automatically. Finally, pronoun resolution will replace personal pronouns, he, she, and they, with their associated proper nouns. This operation is done by the replacement of the pronoun with the most recently used proper noun. For example: "John went to the store. He purchased milk and bread." will become "John went to the store. John purchased milk and bread." The resulting text may sound awkward, but it is still syntactically and semantically identical to the original text, which is why this operation can be done during cleaning and does not have to be done only during the preparation phase.

The syntax-destroying operations of the preparation phase include: removing single letters and noise words, and the replacement of common phrases with its n-gram equivalent. The removal of single letters and noise words is for the aid of the human-in-the-loop, by reducing the number of concepts to be reviewed. The purpose of the D2M process is to reduce the effort required by a human to construct a model, and in the parts to which a human is invaluable we take every effort to reduce the burden. Noise words include articles, prepositions, helping and being verbs. Definition-changing n-grams are identified and words are merged together to form a compound concept. Rather than using n-grams to identify words commonly used together as is used by many information retrieval systems, we use n-grams to identify concepts whose definition changes when taken individually versus part of a compound entity such as "first aid" and "black market". The cleaning phase is done with pregenerated lists of n-grams, noise words, and transformations. Table 3 shows that the number of concepts will actually increase with cleaning and preparation. Recognize that while the deletion of noise words would have a large impact on the number of total words removed, it does not impact the concept count, for example as "the" shows up many

Table 4 Total number of concepts after context sensitive stemming

Data Set	Original	After	Percentage
Sudan	97,492	85,758	87.96 %
Catnet	24,743	22,091	89.28 %
Afghanistan	134,126	129,629	96.64 %

Table 5 Number of nouns/verbs before/after context sensitive stemming

Data Set	Nouns Before	Nouns After	Verbs Before	Verbs After
Sudan	28,488	23,680	12,006	6,763
Catnet	7,693	6,838	4,754	3,223
Afghanistan	71,026	69,317	20,757	20,463

times in the text it is only a single concept and would only reduce the concept count by one. This step is not designed to reduce the number of concepts but to fix the text for subsequent operation.

An additional step, context-sensitive stemming, is done at the end of the cleaning that uses information generated from the syntax-preserved text and uses it for the last step of the cleaning process. A general stemming, such as removing "s" at the end of words, is not done as that would change the part of speech for some words such as many adverbs, and would negatively alter proper nouns. For example, CASOS would be stemmed as CASO, which is incorrect. Stemming, using a Porter stemmer (Porter 1980), is only performed on common nouns and common verbs. The context stemming is provided as a transformation list for human review; however, there are few instances in which the transformation is incorrect. The tradeoff to being correct most of the time is that it will err on the side of not doing a transformation. The context-sensitive stemming has the greatest impact on the normalization of the concepts in the text. Table 4 shows the reduction in the number of concepts. While reduction occurs in all contexts, the extent of the impact will depend on the data set. In contrast to Table 3 in which the deletion of noise words has negligible impact on the total number of concepts that is to be ultimately reviewed by a human, the context sensitive stemming significantly reduces the number of concepts. Table 5 breaks down the reduction further based on nouns and verbs, showing that the amount of reduction varies by both part-of-speech and the data set. Due to the large presence of proper nouns, particularly in news articles, the impact on these data sets is greater on verbs than nouns. Context sensitive stemming is part of the general automated procedure.

The act of stemming is a one of the best examples of where an automated process can significantly reduce the amount of human involvement needed and by which a computer is able to do a more thorough job. Table 6 shows that while the automated approach saves time, it is also more effective than an expert human. We also note that due to cognitive drift, humans tend to alter the way in which they stem as they move through a corpus, thus the results from this automated process are more systematic.

For stemming of proper nouns, we use a near name match operation to build a list of possible normalization merges. This will compare the names and look for slight

Table 6 Impact of different "coders" doing context sensitive stemming on Catnet

Type of Coder	Percentage Reduction	Effort
Novice Human Coder	1 %	20+ hours
Expert Human Coder	11 %	8 hours
Computational Coder	18 %	Less than 3 minutes

spelling variations, of which multiple pluralization and tense fall into that category for proper nouns.

All of these cleaning processes are automated. They are part of the first workflow script. The order and interaction among cleaning procedures has been tuned to capture best cleaning practices.

It is difficult to distinguish a proper noun from a common noun. The part of speech analyzer will more often mistake a common noun to be a proper noun than the other way around, which requires human review. The biggest issue is when a word is at the beginning of a sentence. The proper noun will begin with a capital letter, but so can any common noun found at the beginning of a sentence. It is more unusual for a proper noun to not begin with a capital letter. This approach works well in our favor as it is the case that the deplural/detense list is long, and longer than the names list. In the next phase we will show that the names list will be manually reviewed anyway, in which case it does not significantly increase the burden of manual review by a human. Finally, the part of speech analyzer approach interacts with the machine learning models, based on conditional random fields, that are used to auto-classify concepts into their ontology class as part of the meta-network extraction process.

3.3.3 Step 3: Meta-network extraction

Using advanced text-mining techniques the set of texts are processed to generate the meta-network, the semantic network and an initial project thesauri. Concepts are categorized into their appropriate meta-network ontological class, which includes agents, organizations, locations, events, knowledge, resources, beliefs and tasks (Carley 2002). Agents, organizations and locations are further segmented into specific and generic. The specific concepts, a.k.a. named entities, identify instances of items, such as George W. Bush for agent, UNICEF for organization, and Pittsburgh for location. Some specific entities can be pre-established from existing lists such as countries, major cities, and world leaders.

This process uses the first workflow script. This script first does the cleaning, as just described, and then moves into the meta-network extraction component. This script has optimized the processing to minimize incorrect classification as much as feasible given current technology, and is a fully automated approach to creating a new project thesauri. The detailed decisions that went into the construction of this script are now described. As improved technology is developed it can be "swapped" into the existing script.

Script 1: Choices and process in creating the automated script To create the script that is now script 1 we went through various non-automated and semi-automated processes and identified best practices. These were then ordered correctly for processing

and put into an automated script. This section described the issues faced, choices made, and process of assessment that we went through to create this script. Note, the analyst of D2M benefits from this manual procedure we followed, and need not repeat it.

The first over-arching key issue is degree of generalization. Coding choices can result in the extracted concept set being over- or under-generalized relative to the analysts needs. For example, imagine we are trying to extract all references to a specific person. A low extraction frequency for this actor of interests occurs when the variations of its names, including nicknames and abbreviations, are not properly identified. Misspelling, another source of error, could also lower the extraction frequency. This under-extraction, in turn, results in missing links among actors, which in turn affects the characteristics of the network generated. To avoid this, we found it necessary to include common variation and misspelling of names of the actors, organizations, locations or other named entities of interests in generalization thesaurus, as well as known pseudonyms or aliases. It is also possible to have a falsely higher extraction frequency. This over-extraction occurs, for example, when a generalization thesaurus includes very common nicknames or aliases for a particular actor that are equally applicable to other actors in the corpus. For example, the aliases of "William F. Clinton" in generalization thesaurus should include "Bill Clinton" but should not include only "Bill" as "Bill" is a very common nickname and might apply to other Bill's such Bill Hancock or even Sudan Referendum Bill. If "Bill" were included as an alias of "William F. Clinton," then all "Bill" unidentifiable to any specific actor or entity would then be associated with "William F. Clinton." These false associations would cause "William F. Clinton" to have not only a falsely higher extraction frequency but also false links to other actors, affecting the characteristics of the network generated. We find, in general, for named entities that are n-grams, that single concepts should generally not be associated by generalization thesauri; however, sometimes, advanced machine learning processes can cross-catalog these terms during a second pass (e.g., Diesner and Carley 2008). For non named entities single concepts should be cross classified into knowledge, resources, tasks, and beliefs.

The second overarching key issue is the degree of connection. We found that link extraction was improved by using multiple link-extraction techniques and fusing the results. One technique is proximity based. This technique is widely used in the social sciences for text analysis. Proximity based techniques place a link between two concepts when they occur within some distance of each other; e.g., number of words, sentences, paragraphs. In general, the longer the window (more concepts) the higher the number of links (denser networks); whereas, the smaller the window the fewer the number of links (see Diesner and Carley 2005). The choice of window size is also dependent upon the number of key entities of interest (Bigrigg 2012). A window size of 7 will link concepts within 7 concepts to one another. For example, there are 14 concepts in the previous sentence (the number 7 included). "A", "of", "will", "within", and "to" are eliminated by the standard delete list. "Concepts" merges to "concept" in the stemming process. Thus, the concept "window" will be linked like to: "size", "7", "link", and "concept" but not to "one" and "another." We find that the media impacts this choice of window size; e.g., email, twits, email headers, can be treated with the window size equal to the text, powerpoint slides the window should

be the bullet, and for newspaper data, scholarly articles, and blogs the most accurate connections are at the sentence level. However, this needs to be supplemented with secondary processing to handle lists. Further, it appears for named entities—such as people, organizations and locations specialized paragraph level processing is needed after anaphora resolution is employed.

These issues suggest that for all the texts processed the following should be done:

- Code each text separately and then fused by year.
- Employ n-gram based generalization. Alias thesauri and organizational abbreviations can be constructed from (a) general on-line information on organizations, (b) analyst listing of known aliases, (c) sample estimation from less than 1 % of the texts, and (d) common aliases for world leaders.
- The window size should be set to some predefined level. In general, two sentences is recommended for news articles. That being said, the results in this paper use a window size of seven after anaphora resolution and all generalization thesauri are applied.

We then identified a number of preliminary steps to standardize the extraction process and to eliminate non-bearing concepts from across the texts (Carley 1993). First, we ran standard cleaning processes to clean the data from unwanted characters such as numbers and punctuations. Next, we applied a set of pre-processing thesauri containing pre-made standard and commonly-used delete list in English language (i.e. who, what, the, etc.). We also applied the stemmer from (Porter 1980), a process that converted each concept into its related morpheme to eliminate redundancy of concepts (Jurafsky and Marton 2000: 83, 654). All of these preliminary processing steps are done automatically using AutoMap (Carley et al. 2009a). We tracked the order of processing, experimented with different orders, and identified the best order for reducing error in both what concepts were extracted and the connections among them.

Concept identification is a very involved process that includes deleting uninteresting terms and generalizing others concepts into the set of interest. For example, we transform typos into correct forms, remove plurals, resolve anaphors, delete common stop words, locate known and common n-grams, and we employ thesauri listing common aliases for known political elite, organizations and groups. Abbreviations are also attached to the full concept term in these thesauri. These thesauri are referred to as generalization thesauri. A generalization thesaurus is a two-columned collection that associates a concept with a corresponding higher-level concept (Burkart 2004: 141–154; Klein 1997: 255–261). An example would be associating collie with dog. Thesauri can be used to handle aliases, reduce specific concepts to more general concepts, combine similar concepts, and so on.

In creating the generalization thesauri and identifying the concepts of issues we went through, as noted, multiple iterations. One of the key difficulties, from a cultural inference perspective is handling n-grams. Consider the n-gram Arab League. Clearly all instances of Arab and League cannot be converted to the n-gram. To resolve this issue we created a common n-grams list by (a) processing many documents, (b) looking at wikipedia, and employing subject matter experts (SMEs) to identify specific ones for this region. We found that simply finding all n-grams in the regional corpus, even

only those that occur with a reasonably high frequency, tended to result in a large number of "meaningless" pairs that were simply a function of literary style. Different writing styles result in different "noise" n-grams. SME judgment on subject-oriented ethnography helped to identify relevant n-gram concepts that can be used in the generalization thesauri. Investigation and cross-referencing to the current and historical events was also needed. Thus, generalization thesaurus includes possible variations of the names of the actors, including nicknames, aliases, and abbreviations. Meticulous consultations with Subject Matter Experts (SME) ethnographers, Sudanist SMEs in this case of the Sudan corpus, ensure the accuracy of the final lists and helped identify points where computer assistance could reduce the burden on the SME. Although it is labor-intensive in the front end, this process ensures a systematically standardized and reliable ways to extract information such that it is not prone to errors caused by inconsistencies in human judgment. We operationalized this process by enabling Script 1 to generate these n-grams and use machine learning to auto-classify concepts into categories. Then in the refinement phase, the analyst does this meticulous cleaning just described.

Entity Identification is a theoretically based step; i.e., it depends on the research questions of interests. The complexity of the data processing and analysis increases nonlinearly as the number of entities increases. Thus, there is a trade-off that one needs to consider when deciding the number of entity classes. Since each entity class defines the number of modes, if there is a potential for a minimum of N choose 2 networks that can be extracted where N is the number of entity classes. In our case $N = 2$, and the networks of interest are agent-by-agent (AA), agent-by-organization (AO), and organization-by-organization (OO). This can be further complicated as within each entity class there may be sub-groups. For the Sudan example, in our cases agents are people. These are then further sub-divided into Sudanese and Other. For organizations, these are further sub-divided into ethnic-groups/tribes and other. A second complication is that in texts, entities may be referred to at a generic or specific level; e.g., the president or Bashir. Although we coded for both, in this paper we present only the specific named entities. Script 1 uses machine learning techniques to discriminate generic from specific entities.

Concept classification: This step involves creation of another type of thesaurus, the meta-network or ontology. In general, high quality classification requires employing many classification techniques to create these thesauri. We employ: historical meta-network thesauri developed on other projects including location and world-leader lists, machine learning techniques, parts-of-speech tagging, and specialized meta-network thesauri by SMEs. These meta-network thesauri are used to assign concepts to their corresponding entity class. In this paper, people are identified as agents while groups, including tribes, are identified as organizations. The concept classification in this case is clear as there is generally a clear cut that differentiate entities as either agent or organization. For a more complex ontological scheme with multiple entity classes, this proves to be more difficult as it is possible for concept to be classifiable into more than one entity class (Carley 2002). This problem is rectified by using meta-network attribute thesauri that have more than two columns such that it is possible to identify concepts into multiple classes of entity (Diesner and Carley 2005).

Machine learning techniques are actually quite valuable for ontological classification; e.g., conditional random fields' techniques work well particularly for people,

organizations and locations (Diesner and Carley 2008). Conditional random fields (Diesner and Carley 2008) and other advanced text-mining tools are used to extract additional terms. These tools make use of part-of-speech information and will use the syntax-preserved cleaned text as input though their operations will be applied to the syntax-destroyed cleaned text. Multiple means of categorization are employed, each with their own strengths. A final approach using part-of-speech is employed to be able to provide a category for items not categorized with an indication that this approach is a last guess at a categorization. All verbs will default to task, all proper nouns to agent, all common nouns to resource, and the remaining uncategorized concepts are placed in the knowledge category. The rationale is that humans spend less effort solving true/false questions than multiple choice questions, meaning that during the human involvement the question that the person is attempting to solve is if a concept fits into a particular category. Since the concepts can be grouped by category, the problem is easier to spot the concepts that do not belong. The alternative is to provide no category and allow the analyst to determine the category per concept. D2M, however, uses the CRF approach to create a suggested ontological classification. We note that despite this classification, that for between 5 and 10 % of the concepts, entity classification is dependent on expert knowledge and ethnographic judgment. Future research should determine whether this fraction is a critical fraction or not. Similar to concept identification, this process ensures a systematically standardized and reliable way to extract information such that it is not prone to errors caused by inconsistencies in human judgment. As part of Script 1, not only are named entities extracted, they are classified into the ontological class, and then for agents, locations, and organizations they are further segmented into specific and generic.

Meta-network extraction: This step is very computer-intensive and results in the extracted networks. We use AutoMap to process the set of texts, applying the thesauri, and generate out a meta-network. A large number of procedures are used as part of this process including proximity based mapping, parts-of-speech categorization, and so on. AutoMap searches the whole text set for the concepts previously defined in the generalization thesauri. AutoMap then builds a semantic network based on the generalization thesauri, delete lists, etc., then it cross-classifies the concepts into their ontological categories using the meta-network thesauri, and stores the resulting system as a meta-network. This process causes all the predefined concepts in a text to be linked.

Graph and analyze data: At this point, the text has been turned into a structured data set, specifically a meta-network, that can be graphed and analyzed. We have one such meta-network per text. Since there are thousands of texts there are thousands of networks. The next decision the analyst needs to make is which results to combine. This decision is dependent on the research question. For example, if the issue is how different is the view from the newspaper and the scholars then you would create two fused meta-networks one based on all of those derived from newspapers and second from all of those derived from scholarly writings. Our question about Sudan is somewhat different. We at this point are interested in how the socio-cultural environment in the Sudan has changed over time. Hence, we decided to fuse the data by year. This resulted in one meta-network per year. There are many ways to fuse network data; we chose to employ a unioning procedure. We then took each of the sub-networks by

Table 7 Number of concepts pre-categorized by standard thesauri

Data Set	Before	After	% Reduction
Sudan	114,371	105,294	7.9 %
Afghanistan	121,526	111,934	7.9 %
Catnet	25,907	22,015	15.0 %

year and gave them to the SME's for verification. A yearly compression enabled us to take a longitudinal perspective. This resulted in slight alterations to the thesauri; e.g., in one case we had an organization listed as a person. And it identified weaknesses in the link extraction process; particularly for people to organization links. This was improved with some limited membership inferencing procedures. Overall, however, the SME's confirmed the networks extracted. For graphing and analyzing the data we use ORA (Carley et al. 2011a) as we are dealing with meta-networks over time). Finally, we worked with the SMEs to identify the set of analyses that were always wanted. These included: (a) visualization by year, (b) top-actors (people or organizations) by year on various metrics, and (c) over-all graph-level metrics and change in these.

Our final step was to take all of these findings and solidify them into a workflow that can be used again and on different data sets. We note that while developing the process took months, we can now process that same data even with human-in-the-loop involvement in less than two weeks.

Script 1: Application Through the continued operation of the D2M process, more non-specific concepts may be pre-categorized and its ontology captured into a database to be reapplied to subsequent projects. The current base thesauri (generics and specifics) consist of 150,749 entries, of which 22,455 are generic, are classified as: 784 agents, 321 organizations, 708 resources, 1906 tasks, 1064 knowledge, 1274 locations, and 63 events. Fewer non-specific entries are added as the database of concepts and ontologies grows. The database contains a "noise" list for concepts to be deleted for the convenience of subsequent review. The use of a standard database of predefined categories for concepts reduces the number of concepts that are to be reviewed as shown in Table 7.

Using the pre-established thesauri and delete lists reduces the amount of human involvement by at least 500 %. For example, previously for a corpus of 1109 texts, processing without D2M took over 1000 man hours. Using D2M the AFGHANISTAN assessment took 160 man hours, almost an order of magnitude improvement.

To re-iterate the output of this step is a semantic network, a meta-network, and a project thesauri with both the generalization and the ontology classification components. These are then reviewed by the analyst in the next cleaning step.

3.3.4 Step 4: Refinement (thesauri construction and cleaning)

The refinement step uses both human-in-the-loop cleaning and then application of Script 2. The analyst can repeat the step 3–4 cycle multiple times. D2M supports this process by focusing the analyst on what data needs cleaning and has not been cleaned. Special subtools exist that support the refinement process. These subtools

include tools for: identification of concepts that get converted to a common form, identification of contradictions in thesauri construction, close name identification, and so on.

The refinement step involves:

1. Merging any concepts that need to be merged; e.g., when there are newly identified alias.
2. Deleting concepts that are not of interest to the study.
3. Checking and fixing the classification of concepts into entity classes.
4. Checking and fixing the classification of concepts as specific or generic.

To aid the analyst in this process, information on the concepts and their ontological classification can be provided in a change format file that the analyst can edit. This file displays the following information:

- Frequency
- CurrentConcept
- CurrentMetaOntologyClass
- CurrentMetaOntologyType
- NewConcept
- NewmetaOntologyClass
- NewMetaOntologyType

The analyst can provide a list of changes by simply filling in the new columns if needed. Most analysts will order this list by frequency and simply delete all concepts that occur in only a small number of texts; e.g., in less than ten percent of the texts. This change file can then be used by either ORA or AutoMap to refine the coded data.

To further aid the analyst in this process, information on pairs of concepts that are a near match can be produced. This is useful for identifying simple alternative spellings of the same name. The analyst can provide back the merges from this to either ORA or AutoMap to refine the coded data.

After the first round of cleaning D2M removes all concepts already placed in an ontological category from the review set and provides back to the analyst only the unclassified concepts. Current machine learning tools in D2M are better at classifying named entities, agents, organizations and locations, than other concepts. Consequently, it frequently is more arduous to clean the knowledge, resource, task, event and belief categories than the agent, organization and location categories.

Overtime most analysts will develop two classes of thesauri—universal and project. Universal ones would be used regardless of the project; whereas project ones would be used only on a specific corpi. In general, we find that special terms of art and agent-specific are the only things needed in a project thesauri. Whereas, organizations, locations, agent-generic, knowledge, tasks are often in a universal thesauri. Pre-existing databases, such as terms from wikipedia, or lists of countries from gazateers are often used as part of these universal thesauri. Despite well developed universal thesauri, there will still be occasional terms that are not classified. For example, there are occasional resources not previously identified, such as the name of an item like a Dodge or Chevy referring to a specific manufacturer of a vehicle. These concepts will then need to be classified during the human-in-the-loop phase.

Table 8 Reduction in the number of N grams to be reviewed

Data Set	All 2-3-4-5-n grams	Named entities	% Reduction
Sudan	16,641,044	59,536	99.7 %
Catnet	1,340,802	7,720	99.4 %
Afghanistan	8,891,406	57,640	99.6 %

Script 2 and refinement: choices and process We now describe the issues that we have addressed in designing this part of the D2M, and describe manual approaches that were initially taken. As in step 2, the analyst who uses the D2M does not need to go through these same manual processes as we basically automated as many as feasible.

One of the first choices made had to do with locations. While locations of interest, such as cities and countries, are provided by a location database, a gazetteer, some locations are for place names not commonly found in a gazetteer, such as neighborhoods or street names. Location databases will include this information, but we do not include them as these do not show up frequently and to use them generally would create many false hits. We note that often place names and agent names are the same. In addition, common terms such as "But" are also location references.

A second choice is centered around how to treat n-grams as a single concept. As noted, project-based generalization is performed. The generalization process serves many purposes: n-grams, normalizations (Jurafsky and Marton 2000: 83, 654), and replacement. The n-grams will merge multiple words into a single concept. Text analysis operates on the level of a "token" meaning a unit of text that is distinguished from the rest of the text. Typically a token is a collection of letters (a word) demarked from the rest of the text by a leading and training blank space. There are instances in which a single word does not represent the entire concept being described, in the case of specific entities such as names of people, e.g. George Bush, or organizations, e.g. United Nations, or locations, e.g. New York. In AutoMap, the use of an underscore between related words is used to denote that the individual words should be considered as a single concept.

A third choice centered on the provision of n-grams to the analyst for review. We experimented with providing all possible n-grams en masse for review. The approach was to extract all possible 2- 3- 4- and 5- word sequences, along with their frequency of occurrence, and provide that list to a human for review. The list was prohibitively long with a low signal to noise ratio. Hence, in D2M only bi-grams are automatically provided and classified.

Given that one of the primary types of n-grams is the multiple words that make up a name of a person, organization, or location, n-grams are not formed from all possible word sequences, but from all sequences of proper nouns. This part-of-speech analysis is used to identify proper nouns as so-called named-entities. This has cut down the list of possible n-grams significantly. The adjacent proper nouns are listed together as n-grams as many project-based specifics are compound concepts, especially names. This reduces the number of n-grams to review to a manageable set, as shown in Table 8.

The close concept approximation procedure described earlier that aids in stemming for proper nouns, is applied to the names to find alternative spellings. The list

Table 9 Close approximations

Study	Number of Close Approximations
Sudan	795
Catnet	125
Afghanistan	44

of alternative spellings and proper noun stemming is small, as shown in Table 9, but the routine will find these items easily as they are difficult to identify by the human. The list is provided to the human for review. Note, this approach will identify names differing in a few letters such as Mohamad and Muhamad but will not identify cases where first and last name are place in reverse order such as Barack_Obama and Obama_Barack. Generally, the analyst can quickly identify whether the two "close" names are the same, or not. The analyst marks defines which form is desired for the way the concept will appear in ORA and new thesauri entries are created. These are then stored and become part of the permanent thesauri for that project.

The D2M process will extract a meta-network (nee social network) as it appears in the text (Diesner and Carley 2005). Concepts that do not contribute to the final meta-network can be removed early in the text processing, and then will not appear in future codings. Note, AutoMap can be used to extract a semantic network to review the uni-modal network relationships between concepts, and in such a case the noise words would not be removed. In the D2M process, this semantic network is generated, such that, all of the concepts deleted and all of the thesauri conversions are already done.

A key part of the refinement process is reviewing the list of concepts. A typical human approach to reviewing lists of concepts is to sort the list based on the frequency of occurrence of the concept in the text corpus. D2M facilitates this by providing that information in the change file. We note that this approach does bias the results toward the most frequent concepts. If importance is correlated with frequency, then this bias is not necessarily problematic. However, commonly the most frequent possible n-gram is "of the" followed by "in the" and "to be". Further, common concepts are often low content bearing words such as "the." It is recommended that the analyst delete these concepts from the corpus. ORA has a subroutine for determining the linguistic reach and so importance of concepts—this can be used to filter concepts of low reach. This report in ORA is the semantic network approach.

Script 2: Deconfliction The project-based thesauri and delete lists capture differences in context-based word usage: e.g., a military and a weather "front". They are constructed using a computationally supported human-in-the-loop workflow. The role of the human is to review the generated thesauri for alternative categorization of concepts.

To reconcile categorizations done by multiple tools, and by multiple iterations of review, we have developed a merge tool that will combine multiple thesauri into a single final reconciled thesauri, using the idea that the entries in one thesauri will take precedence over the entries in another thesauri. The two thesauri are merged with a higher precedence thesauri being applied to the lower precedence thesauri, allowing the higher precedence thesauri entries to take precedence. For instance, if

a concept is categorized as a resource, such as in the part-of-speech approach, yet a better approach using machine learning approach categorizes the same concept as an organization, the final category for the concept would be organization. The merge tool will also remove unnecessary transformations such as A being transformed into B which is transformed into C. The merge tool would replace the structure with A being transformed directly into C which will improve processing time.

Using the two column format for a generalization thesauri would perform a find-and-replace operation replacing all instances of one word or phrase to an associated word or phrase (Burkart 2004: 141–154; Klein 1997: 255–261). The find-and-replace operation is one component of the master thesauri. The final concept is then linked to its ontology and to its associated attributes. The data that otherwise would exist in multiple files is consolidated into a single file for the convenience of the analyst, whereas its original storage may be in a structured database.

The construction of a master thesauri is another advantage of the D2M process. The master thesauri is a file or database that contains concept transformations and removal, ontology categorization of concepts, and attributes associated with concepts. The thesauri format is such that it encourages the relationship between an original word/phrase, the resulting concept, its ontology, and its attributes. The single file format reduces relationship errors such as:

File 1: United States,United_States
File 1: United States of America,United_States
File 2: United_States_of_America,location

It is difficult for a analyst to conceptualize the final concept, which is United_States in File 1, yet was assumed to be United_States_of_America in File 2. Each of these final concepts is an equally suitable choice, provided only one is chosen as the representative of the concept and used consistently.

All meta-networks, with or without augmentation, are stored in DyNetML. DyNetML is the interchange language that supports the D2M process flow work.

3.3.5 Step 5: Final meta-network extraction

As noted, the analyst will repeat the processing step 3 using the new thesauri and delete lists of step 4. This sequence is repeated until the analyst is satisfied; typically 3–5 processing rounds are sufficient.

Once the analyst is satisfied with the refinement of the concepts the final meta-network is extracted. This is a completely automated process. All thesauri and coding choices selected earlier are applied. The resultant network can then be read into ORA for assessment or it can be augmented with additional data.

It is important to note that meta-network will include both the semantic network—as a network of all concepts to all concepts and a set of subnetworks based on the ontologically cross-classified concepts. This meta-network will include both forms of the cross-class networks. A cross-class network is a bipartite network; i.e., a network formed from two entity classes, such as agent x organization and organization x agent. The analyst may choose to run the reduced form sub-routine that combines these two bi-partite into a single network by inverting one and then adding the networks. In addition, many analysts remove the semantic network or put it in a separate

meta-network for processing. Finally, some analysts prefer working only with specific concepts and so use the remove nodes procedure in ORA to remove all generic agents, generic organizations and generic locations.

3.3.6 Step 6: Data augmentation

Data augmentation uses additional data sets and rules of inference to augment the extracted networks. Three types of additional data are typically used: geo-spatial, attribute files and meta-data.

Geo-Spatial Augmentation: Using post-processing geo-information, latitude and longitude, are added to facilitate analysis and visualization. For this purpose we used GeoNames (http://www.geonames.org), which is an open source gazetteer site. Alternatively NGIA GeoNET (http://earth-info.nga.mil/gns/html/) could be used. These sites contain geospatial-information, and other data, such as population and location type (city, village, etc.) that can be added as attributes through a human-in-the-loop process.

Attribute files augmentation: Attributes are information about nodes that are of routine important. Note specific and generic are an example of an attribute. Other types of attributes may be items like GNP for countries or gender for agents. These might exist in an excel spread sheet or flat CSV file. They can be read in by ORA and used to augment the existing data. By including database information early in the process, the results provided by the analyst are constrained to be results that can be easier to link back to the database such as attribute information. Attribute information can be simple values such as latitude and longitudinal properties of a location, or can be complex information about a person. If the reconciliation of the concept as it appears in the text and the concept as it appears in the database is left until later in the process, an additional manual process of reconciling concepts has to be done.

Meta-data: Meta-data is that information in the structured part of a semi-structured data. For example, in email, this is the header information. Postprocessors exist for certain types of meta-data for adding them to the AutoMap produced meta-networks.

The final approach to augmentation is inference. Using findings from anthropology and organization science, ORA has a series of procedures for inferring beliefs and group membership. These can be applied to the meta-network resulting in new nodes and links. The improved meta-network can then be saved and processed.

During step 3, refinement, all data in a corpus was processed as a single network containing all data per project. Once cleaning is done, the data is typically split into several networks, each produced using the same thesauri. This splitting or binning process typically reflects geographic, temporal or source differences that are critical in the analysis. For example, in the CATNET study the data is binned by source—news-site or website. Whereas, in SUDAN the data is binned by year, with one network produced for each of the 10 years. In all cases, it is neither necessary nor advisable to develop a different thesauri for each bin as the single thesauri contains the information to be applied to all networks. This supports the analyst's ability to compare and contrast socio-cultural behavior across the data.

ORA has features in it for comparing and contrasting the meta-network before and after augmentation. To date, augmentation is a mixed-initiative process. More research is needed to determine which aspects of this process can be further automated.

Table 10 Resulting concept by ontology count

Ontology Class	Catnet	Afghanistan	Sudan
Agents	2874	1216	996
Locations	1287	2781	2671
Organizations	635	1225	487
Task	1036	638	1231
Resources	364	251	310
Event	189	41	52
Knowledge	407	265	991
Beliefs	14	48	1
Concepts	22,516	106,206	6,739

3.3.7 Step 7: Network analytics and forecasting

The resulting meta-network can then be processed using network analysis routines in ORA (Carley et al. 2011a). For example, the analyst may want to characterize the meta-network model in terms of the number of various entities, or what entities are key in the network. Table 10 shows the aggregate counts of the number of concepts per ontology as calculated via ORA. Note that the number of concepts per category is dependent on the data set used, regardless of the number of entries in their project thesauri. All possible network analytics can be run on the resulting meta-networks.

Using ORA key actors, resources and hot-topics are identified thus defining possible targeting and intelligence operations, or communication interventions, and humanitarian assistance events. Illustrative key actors and beliefs are: AFGHANISTAN: Mahmud Ahmad, nuclear action needed; CATNET: Omar Bakri, influence over time; SUDAN: Minnawhi, separation of S. Sudan. These interventions based on the key nodes in one or more entity class can then be used to set up a series of virtual experiments which can be assessed using a dynamic-network agent-based simulation.

The human analyst selects the interventions of interest and then using the experimental design system builds a virtual experiment to examine the impacts. We use Construct (Carley et al. 2009) for this process. Construct is embedded in ORA as part of the Near Term Analysis Process. The analyst might at this stage assess the impact of removing critical nodes, links, or adding nodes or links. Across all three data sets, Construct can be used to identify changes in the social network, the knowledge network, the distribution of resources, information and beliefs. Critical actions such as the removal of key nodes can be assessed. For example, for the data sets used here, removal of the identified key actor results in strengthening of the belief associated with the hot topic.

4 Summary

D2M involves multiple workflows, i.e. sequences of steps, applied to diverse data. Many of these workflows have been automated thus freeing the analyst to focus on the domain. By sharing workflows across these domains a consistent approach was

established. As advances in the underlying technologies are made, the workflows in D2M can be easily adapted and D2M rerun. We find that D2M reduces analysis time from years to less than a month, or two-weeks when using parallel processing. The results are more systematic. The strain on the analyst is less. And, overall coding accuracy is improved.

Several challenges exist, whose resolution will improve D2M. Improvements are needed in geo-information extraction, interpreting non-English embedded concepts, automated re-merging of email/chat content and header data, automated removal of meta-data, auto-topic identification to supplement analyst defined topic identification, post-processing for event and sentiment extraction, semi-automated simulation instantiation, and support for constructing virtual experiments. Current efforts are underway to improve the context sensitive stemming to support improved depluralization of proper nouns, and conversion of all tenses of a verb into a common verb.

The refinement process currently uses concept frequency as a key filter. Future research should explore whether detailed cross classification of low frequency concepts actually impacts the final results and in what way.

Despite these challenges, use of D2M in its current form, makes possible the rapid assessment of vast quantities of data as networks. Using D2M the analyst can identify not just trends in what is talked about, but changes in the underlying network structure of the community being talked about, identify key leaders, and identify changes in who is talking about what.

Acknowledgements This work was supported in part by the Office of Naval Research—ONR-N000140811223. (SORASCS), ONR-N000140910667 (CATNET), ONR-N000140811186 (Ethnographic), and W15P7T-09-C-8324 awarded by CERDEC-C2D under the THINK ATO. 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, CERDEC, the Department of Defense or the U.S. government.

References

Alexa M (1997) Computer-assisted text analysis methodology in the social sciences. ZUMA-Arbeitsbericht 97/07

Batagelj V, Mrvary A, Zaveršnik M (2002) Network analysis of texts. In: Erjavec T, Gros J (eds) Proceedings of the 5th international multi-conference information society—language technologies, Ljubljana, Jezikovne tehnologije/Language Technologies

Bauersfeld K, Halgren S (1996) "You've got three days!" Case studies in field techniques for the time-challenged. In: Wixon D, Ramey J (eds) Field methods casebook for software design. Wiley, New York

Beebe J (1995) Basic concepts and techniques of rapid appraisal. Human Organ 54(1):42–51

Bentley ME, Pelto GH, Straus WL, Schumann DA, Adegbola C, de la Pena E, Oni GA, Brown KH, Huffman SL (1988) Rapid ethnographic assessment: applications in a diarrhea management program. Soc Sci Med 27(1):107–116

Bigrigg M (2012) Window size effect on key network entities. Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report, CMU-ISR-12-102

Blei DM, Ng AY, Jordan MI (2004) Latent dirichlet allocation. J Mach Learn Res 3:993–1022

Burkart M (2004) Thesaurus. In: Kuhlen R, Seeger T, Strauch D (eds) Grundlagen der Praktischen Information und Dokumentation: ein Handbuch zur Einführung in die Fachliche Informationswissenschaft und -praxis. Saur, Munich

Carley KM (1993) Coding choices for textual analysis: a comparison of content analysis and map analysis. Sociol Method 23:75–126

Carley KM (1997) Network text analysis: the network position of concepts. In: Roberts CW (ed) Text analysis for the social sciences. Lawrence Erlbaum, Mahwah

Carley KM (2002) Smart agents and organizations of the future. In: Lievrouw L, Livingstone S (eds) The handbook of new media. Sage, Thousand Oaks, pp 206–220

Carley KM (2006) Destabilization of covert networks. Comput Math Organ Theory 12:51–66

Carley KM, Martin MK, Hirshman B (2009) The etiology of social change. Top Cogn Sci 1(4):621–650

Carley KM, Reminga J, Storrick J, Columbus D (2011a) ORA user's guide 2011. Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report, CMU-ISR-11-107

Carley KM, Columbus D, Bigrigg M, Kunkel F (2011b) AutoMap user's guide 2011. Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report, CMU-ISR-11-108

Chakrabarti S (2002) Mining the web: analysis of hypertext and semi structured data. Morgan Kaufmann, San Mateo

Corman SR, Kuhn T, McPhee RD, Dooley KJ (2002) Studying complex discursive systems: centering resonance analysis of communication. Human Commun 28:157–206

Diesner J, Carley KM (2005) Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis In: Causal mapping for information systems and technology research: approaches, advances, and illustrations. Idea Group Publishing, Harrisburg

Diesner J, Carley KM (2008) Conditional random fields for entity extraction and ontological text coding. J Comput Math Organ Theory 13:248–262

Ding B, Zhao B, Lin CX, Han J, Zhai C (2010) TopCells: keyword-based search of top-k aggregated documents in text cube. In: Proc of 2010 int conf on data engineering (ICDE'10)

Garlan D, Carley KM, Schmerl B, Bigrigg M, Celiku O (2009) Using service-oriented architectures for socio-cultural analysis. In: Proceedings of the 21st international conference on software engineering and knowledge engineering (SEKE2009), Boston, USA

Hofmann T (1999) Probabilistic latent semantic analysis. In: Proc of uncertainty in artificial intelligence

Holsti OR (1969) Content analysis for the social sciences and humanities. Addison-Wesley, Reading

Jurafsky D, Marton JH (2000) Speech and language processing. Prentice-Hall, Upper Saddle River

Klein H (1997) Classification of text analysis software. In: Klar R, Opitz O (eds) Classification and knowledge organization: proceedings of the 20th annual conference of the gesellschaft für klassifikation eV University of Freiburg, Berlin. Springer, New York

Krackhardt D, Carley KM (1998) A PCANS model of structure in organization. In: Proceedings of the 1998 international symposium on command and control research and technology evidence based research, Vienna, VA, pp 113–119

Krippendorff K (2004) Content analysis: an introduction to its methodology, 2nd edn. Sage, Thousand Oaks

Landauer T, Foltz PW, Laham D (1998) Introduction to latent semantic analysis. Discourse Process 25:259–284

Lin CX, Zhao B, Mei Q, Han J (2010) A statistical model for popular event tracking in social communities. In: Proc of 2010 ACM int conf on knowledge discovery and data mining (KDD'10)

Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press, Cambridge

Popping R (2000) Computer-assisted text analysis. Sage, Thousand Oaks

Porter MF (1980) An algorithm for suffix stripping. Program 14:130–137

Ramakrishnan C, Kochut KJ, Sheth AP (2006) A framework for schema-driven relationship discovery from unstructured text. In: Proc international semantic web conference

Roth D, Yih W (2007) Global inference for entity and relation identification via a linear programming formulation. In: Getoor L L, Taskar B (eds) Introduction to statistical relational learning. MIT Press, Cambridge

Wang C, Han J, Jia Y, Tang J, Zhang D, Yu Y, Guo J (2010) Mining advisor-advisee relationships from research publication networks. In: Proc 2010 ACM SIGKDD conf on knowledge discovery and data mining (KDD'10)

Zhang D, Zhai CX, Han J, Srivastava A, Oza N (2009) Topic modeling for OLAP on multidimensional text databases: topic cube and its applications. Stat Anal Data Min, 2:378–395

Kathleen M. Carley is a Professor of Computation, Organizations and Society in the Institute for Software Research Department in the School of Computer Science at Carnegie Mellon University and the director of the center for Computational Analysis of Social and Organizational Systems. She received her Ph.D. in Sociology from Harvard and her undergraduate degrees from the Massachusetts Institute of Technology. Her research interests are in dynamic network analysis, agent-based modeling, social networks, and information diffusion and belief dispersion. In addresses her interests she takes an interdisciplinary approach drawing on advances in machine learning, organization science, social-psychology and cognitive science. She and her team developed ORA, AutoMap and Construct.

Michael W. Bigrigg is a Project Scientist with the CASOS Center in the Institute for Software Research at Carnegie Mellon University. He received a Ph.D. in Computer-Aided Engineering from Carnegie Mellon University, a M.S. in Computer Science from the University of Pittsburgh, and a B.S. in Computer Science from Indiana University of Pennsylvania. His research interest is at the intersection of social, communication, and information networks.

Boubacar Diallo a Fulbright Alumni from Burkina Faso (West Africa). He graduated from Indiana University of Pennsylvania with a Bachelor degree In English and with honor Cum Laude. While working on this paper he was a technical writer with the CASOS Center in the Institute for Software Research at Carnegie Mellon University. His research interests include education and politics.