Digging into Human Rights Violations: Data modelling and collective memory

Ben Miller*, Ayush Shrestha*, Jason Derby*, Jennifer Olive*, Karthikeyan Umapathy†, Fuxin Li‡, Yanjun Zhao*
*Georgia State University
miller,ashrestha2,jderby1,jolive1,yzhao9@gsu.edu
†University of North Florida
k.umapathy@unf.edu
‡Georgia Institute of Technology
fli@cc.gatech.edu

***Abstract*—Archives of human rights violations reports, by virtue of their poor metadata, basis in natural language, and scale, obscure fine grain analyses of violation event patterns. Cross-document coreference of victim or perpetrator occurrences from across a corpus is challenging, particularly when those mentions relate to different events. These challenges are emblematic of the transition from small scale to big data analysis in the humanities. This paper discusses these issues and proposes a framework to address these challenges so as to explore narrative construction and the formation of collective memory. Though our framework is based on processing human rights violation reports, it can be readily extended to support other big data problems in the humanities.**

***Keywords*-Digital Humanities; Big Data**

I. Introduction

Records pertaining to human rights violations are principally consulted so as to better understand the history of an event or potential responses to ongoing events. These records have heterogenous origins, capturing material produced by civilians, governments, and NGOs, and by victims, observers, and perpetrators. They are produced both during an event, frequently as governments bureaucracies document their own behavior, and after an event, as witnesses emerge to speak out against violators, or to participate in truth and reconciliation proceedings. They contain heterogeneous information types ranging from aggregate lists of atomized acts to geospatial information regarding sites significant to the prosecution of violations such as mass graves to interview or interrogation transcripts to observation reports by professional observers. The desired analytic outcomes are equally broad, encompassing, 1) attempts to quantify the scope or frequency of violations so as to make determinations of the character of a violation pattern, 2) determine emerging patterns of violations and assess possible interventions, 3) attempts to study the generalizability of a given records collection in relation to a violation context, 4) attempts to gather correlated evidence for truth and reconciliation or prosecutorial efforts, or 5) attempts to tell the history of an event, for the assuaging of public memory, for the scholarly record, or for the prosecution of suspected violators. This heterogeneity of form, a heterogeneity not uncommon for big data problems in the humanities, makes analysis challenging. As one example, consider that most corpus processing methods only function for narrowly de fined data models: e.g. a key words in context script in R relies on the predictable extraction of well-transcribed text.

Although human rights corpora are smaller than the datasets addressed as big data in the sciences or social scientific projects examining the social web, various features of human rights data and its analysis make this a big data problem. These features include the high dimensionality of this information, the heterogeneity and number of reports in a given corpus, the population level coverage of the corpora, the requirement for real-time analysis, and the analysis requirement for veridicality. Methods that address these requirements are drawn from the core elements of big data analysis: statistics, machine learning, data visualization, data architectures, corpus linguistics, high performance com puting, and HCI.

Survivors and witnesses to human rights violations are placed in the complicated and problematic place of describ ing and relating information about a traumatic event. In each persons individual recall, incidents and details regarding the event may not always line up with one another perfectly. This is problematic as it creates multiple accounts of a singular event; however, it is also interesting in that it allows for other considerations to be made to a narrative history that make up a collective memory and shape cultural identity regarding the traumatic event. Problems, however, abound when considering the scope of human rights violations as many narratives become part of the larger narrative and not always in the same form. How does one create meaning from these records collections when their scale and fragmentary nature resist access and interpretation? How does one reveal meaning in ways that preserves the specific language and character of the witness observations? How can NLP assist humanistic researchers in the performance of these tasks? From photos to memos to transcripts to field reports, portions of traumatic narratives live in a variety of documents that are not only multiplied by the number of atrocities, but also by the ways in which they are recounted. To this end, big data analysis can serve as a tool for constructing a narrative

978-1-4799-1293-3/13/$31.00 ©2013 IEEE

out of many.

Our framework is designed to process large numbers of narratives, parse elements describing time, location, person, and semantic context, and allow for larger narratives to emerge. This cross-document narrative can then be used to understand how collective memories are formed and how history is made. This same process can also identify perpetrators and victims on the basis of their linguistic and narratological structures, and to contrast how groups self identify, on linguistic levels. This transversal reading of large text archives of human rights abuses can also facilitate the discovery of the stories of victims and perpetrators hidden by the scale of these corpora.

II. Related Work

A. Human Rights Data Analysis

Narratives pertaining to the violation of human rights have occupied the humanities since the development of linear historiography [1] with the Homeric retelling of the annihilation of Troy, and documentation at scale has existed since the Federal Work Progress Administration sent writers throughout the American South to interview 2,300 former slaves between the years of 1936 and 1938. What differentiates the contemporary context is twofold. First, the scale of the records collections that document atrocities has grown astronomically, and second, the records being analyzed are in some instances produced contemporaneously with the violation. Three examples indicate different aspects of the scale of contemporary data analysis in human rights violations research: 1) the Guatemalan National Police Historic Archive is estimated to contain approximately 80,000,000 text records, 2) two organizations, the Shoah Foundation and the Fortunoff Archive for Holocaust Video Testimony have combined to produce over 100,000 hours of Holocaust survivor testimony, and 3) in trying to estimate the number of casualties in the current Syrian crisis, the Human Rights Data Analysis Group (HRDAG) had to work across the ledgers of eight separate organizations, each of which was conducting idiosyncratic casualty counts over varying periods of the conflict. Near real-time reporting of violations both enables human rights workers to respond to events in progress, such as in the work of Best et. al. on the use of mobile devices in election monitoring [2], and the work of Norheim-Hagtun and Meier on real-time crisis mapping in the aftermath of the 2009 Haiti earthquake [3].

In order for scholars from the humanities to engage with emerging big data analytical tools important changes are required to the weakly structured nature of traditionally collected data. These changes both enable and proscribe the analyses that scholars can perform upon their data sets. Migrating data, for humanities scholars, relied traditionally on methods of sampling, such as qualitative coding, string matching, or transposition. These classifications and operations limit data to normative values. Data storage technologies such as structured databases also limited humanistic data, and required careful processing of data sources and schema design to ensure informational integrity. One way to redress the necessity of sampling operations could be derived from expanding data storage techniques associated with big data such as non relational database systems. These possibilities present tempting options for humanistic inquiry as they can reduce the necessity to preprocess data via sampling methods.

This project is predicated on the premise that given a large collections of records describing a population, figures will recur in multiple documents. The idea that records over-represent victims in explicit and implicit ways emerges from research by Silva and Klingner et. al. in [4], and Patrick Balls 2003 report on the number of killed or missing Peruvians from 1980-2000 [5]. That research examines collections of human rights documentation, such as the 49,000 documents recovered from the Chadian Documentation and Security Directorate, and the 24,000 reports of missing Peruvians submitted to the Comisin de la Verdad y Reconciliacin (CVR). By applying statistical methods for population counts such as Multiple-Systems Estimation, [5], [6] have screened collections of reports about victims of abuse for recurring individuals, and provided estimates of how many people enumerations missed. That relatively simple statistical models of population analysis can determine quantitatively that victims of violence show up in multiple documents within a collection implies that other methods can be used to determine who those victims are. Their methods, by necessity, focused not on identifying those individuals, but on using statistical modelling of archives to determine the number of victims. This over-reporting of individual victims within human rights corpora is endemic, and presents an opportunity for a system that can read across a corpus. Our work builds on their quantitative approach to human rights records by offering a qualitative method for assembling the fragmented stories of those victims. In our method, data provides the framework within which human beings communicate meaningful testimony; that testimony is embedded in, distributed through, and obfuscated by the archives of human rights violation reports. The difficult work this project takes on is identifying the recurrence of those individuals and extracting the text that embodies that recurrence.

Our first argument, that fragmentary descriptions of perpetrators, victims, and abuses recur throughout a collection is important in human rights and truth and reconciliation contexts. Revealing these implicit stories adds evidence to truth and reconciliation efforts, and supplements historians access to the stories of an event. This requires identifying related fragments of text within documents from across an archival collection, and stitching them together in ways that preserve context to present a coherent, accurate figure. This mode of reading resembles traditional documentary

biography, and is already challenging with limited corpora, such as a victims diary or a collection of reports from one police precinct. When facing archives exceeding millions of words, reading is impracticable, and marginal figures effectively invisible.

B. Big Data Models

The choice of data model is complicated as the big data turn implicates data collection methods, both enables and constrains analyses, and influences the constitution of knowledge [7]. Replicating data models more comfortable with high-dimension data such as free text may expand the analytical possibilities available to humanities scholars. For example, big data analytical tools such as those developed by the Culturomics project [8]or cultural analytics project by L Manvoich [9] require that researchers first render texts into accessible forms. Both projects narrow the analytic scope to features susceptible to computational methods such as saturation or simple strings. In order to leverage the power of big data tools, researchers need to develop models that balance scalability, flexibility, and mobility.

C. Information Modeling and Extraction

Most commonly, approaches to information extraction in free text blend low context attempts at statistical analysis of material restricting scope to an internal frame of reference, and high context approaches that seek to link extracted information to external data stores such as Wikipedia [10]. A low context approach for a corpus study of human rights violations reports might see the problem as one of document clustering and apply a method such as Term Frequency - Inverse Document Frequency so as to build identifying keyword sets for each document. Those keyword sets, when compared, imply document families. A more complex, but ideologically similar approach, might use a technique from latent semantic analysis to infer topics for each document, and produce clusters based not on explicit keywords but on implicit topics. These techniques look to model a corpus based on typicallity at the level of the document, and thereby facilitate a human researcher's engagement with a collection through categorization at a particular level of granularity. A high context approach would check extracted information against an external reference so as to attempt more complex tasks such as disambiguation, or geocoding of place names so as to map the various locations in a report. Low context approaches seem best at the type of general engagement with a corpus described as distant reading [11], while high context approaches attempt to synthesize information beyond the limit of a given record. Other information extraction and emplotment projects such as ChartEx [12] and Trading Consequences [13] demonstrate two approaches to the problem of disambiguation of entities and place names. The first relies on a diagrammatic approach to locations and rigid genealogies of entities. The second correlates extracted location names to mapped places, but then has to do so at a very general, country level at which disambiguations do not occur.

Research into the problem of Information Extraction (IE) from unstructured and fragmentary documents has yielded many techniques, but few are usable by the non-computer science researcher, and few work for the particularities of large collections of violations reports. Often, the report formats are idiosyncratic, metadata is sparse, English is frequently a second language, and narrative is the dominant conceptual mode. These reports begin as free text that quickly gets sublimated in a reporting framework so that the violation information can be separated from the stories. Soderland et al. [14] describe many computational techniques in the area of IE that have been applied to various domains such as game scores from the National Football League, and the Intelligence Community. The difficulty of using these techniques and their inherent limits in the face of unstructured text has left witness reports, medical narratives, and many other domains unaddressed.

Humanistic work to elicit narrative connections across sets of documents pertaining to traumatic events has been limited to visualizations of the metadata contextualizing the free text, and statistical understanding of populations and casualties. These elisions were dictated by the purpose of the projects and the difficulty presented by unstructured text. For example, visual explorations of the Afghanistan War Logs by McCormick, et al. [15] for *The Guardian UK* focused on geographic, abstract understanding of the Logs metadata: stories behind the data were necessarily elided.

Work in the broader field of Narrative Intelligence of pattern detection and representation of a series of linked acts was pursued by an interdisciplinary reading group at MITs Media Lab in the early 1990s [16]. Frameworks to facilitate the development of textual data analytics tools include UIMA, GATE, Seeker and SemTag, Cerno, and Armadillo. Like the more familiar SemTag, Cerno [17] and Armadillo are focused on Semantic Web applications and so are most useful for automated tagging and retagging of documents in a corpus. Each produces XML annotations of the source text, and definitions for the annotations. Of the three, Armadillo is the most flexible in regards to document regularity, and the most automated. UIMA and GATE are extensive in terms of the text processing modules and are open source, allowing for further customization and extension. Each allows for analysis of document corpora through the tokenization and parsing of the text as symbolic and syntactic objects. Additional functionality is available in both frameworks for pipelining Java programs that introduce AI text-mining methods such as LSA, or for structured regular expression searches based on user definitions. Methods that rely on user-defined patterns are both time consuming for the end user and fundamentally miss the central problem presented by implicit figures in large text archives the

patterns are not visible to unaided readers because the scale of the archive resists conventional modes of reading. Armadillo relieves some of the burden on the end user but retains a focus on highly local identifications, i.e., to what category does a particular phrase belong. Despite the wealth of computational techniques for text analytics, a gap exists between a humanistic researcher's conceptualization of the tasks required to make sense of records and the computational models currently used by IE tools.

D. Anaphora Resolution

Anaphora resolution has been a topic of interest in NLP since the work of Sidner[18] on pronominal anaphora resolution using domain and linguistic approaches, and Hawkins [19] on associative anaphora. The period from 1990 until today has focused on shallow, low context, syntactic approaches [20][21]. Anaphora, most often recognized in pronouns, refer to any substitution of an ambiguous phrase for a more specific phrase. Similarly, exophora occur when the specific subject is not present in the current document. The current state of the art for computational resolution of anaphora is best described in the work of R. Mitkov, S. Lappin, B. Webber[22], M. Denber, and M. Dimitrov[23]. Anaphora resolution systems have been developed and examined in the work of Mitkov, Webber, Lappin and others, and are classified as knowledge-poor or knowledge-rich. An example of a knowledge-poor system, i.e. one that does not take into account complex linguistic, semantic, or grammatical rules, is Dimitrov et als GATE-based implementation. Dimitrovs system drew on earlier work by Lappin and Leass [24], [25] pronominal resolution that used indicators such as definiteness, heading, collocation, referential distance, and term preference. Mitkovs system achieved pronominal resolution rates of 89.7% within a corpus of technical manuals[23]. Dimitrovs system read a corpus of approximately 180,000 words drawn from broadcast and print news sources. These genre specifications and corpora limits are typical of anaphora resolution systems. Our work began with the core of Dimitrovs knowledge-poor system.

III. Method

Our framework describes a low context approach to the problem of transversal reading. We propose a two tiered framework with Data and Presentation layers. The data layer is responsible for data extraction, parsing and running NLP modules to get the entities and events. The presentation layer handles the visualization of the data and the feedback loop which modifies the data based on the user feedbacks.

A. Data Layer

This layer deals with how the data is extracted, processed, and stored in the backend, supporting the presentation layer. This layer consists of the extraction module and the NLP module. The extraction module is responsible for digitizing and parsing the documents. A large portion of human rights violation documents that we obtained were stored in hard copy format and, in many cases, handwritten. We used commercial OCR tools like OmniPage to digitize some of the documents. In cases where the materials are barely readable due to poor archiving, manual transcription is required. After digitizing, the document is parsed and the output is routed to the NLP module.

The goal of the NLP module is to facilitate cross-document coreference of entities in collections of witness statements and interviews within the domain of rights. The module's approach to resolving the task [26] described as "whether or not two mentions of entities refer to the same person," begins by considering exophora and relies on placing pronominal entities within a high-order Event Storygram of location, time, name, and semantic context. Because temporal information is so often referential and ambiguous, and therefore difficult to extract and correlate [27] our approach uses a phrase-based establishment of semantic context to support identifying the temporal context and to reinforce the automatic matching of elements within the Storygram. Our model for processing linguistic uncertainty extends the work on veridicality in [28], [29], [30], incorporates semi-supervised machine learning methods with a taxonomy. This model describes the validity of automatically generated correspondences amongst the relations of person to time to location to semantic context.

This noun- and verb-phrase extraction, collocation detection, and semi-automated matching, feeds a 2D-planar visualization similar to network graph models. Uncertainties in the document and information retrieval processes are visualized to allow researchers to confirm whether entity occurrences should be conflated. Because much human rights documentation contains sensitive information that cannot be made public, this project is prototyping with both publicly available and restricted data. Publicly available data sets used in this work include interviews with first responders to the World Trade Center attacks and documentation exproduced by or related to the Extraordinary Chambers in the Courts of Cambodia (ECCC) as well as redacted version of reports describing contemporary violations committed by the Lord's Resistance Army in central Africa.

1) Event summarization based on matching phrases: Our main stratagem is to situate entities in the series of events that define their appearances. Phrases useful for this process accord to a "journalist template," of Who, What, When, Where, and Why, and are situated in the events reporting schema developed by Patrick Ball for human rights violations reporting [5], [6]. The goal is a system that can automatically extract these important entities as phrases and based on these extracts, allow for the recognition of duplicate entities across documents. A perceptual diagram of the system is shown in Figure 1.

After the noun and verb phrases are extracted and passed

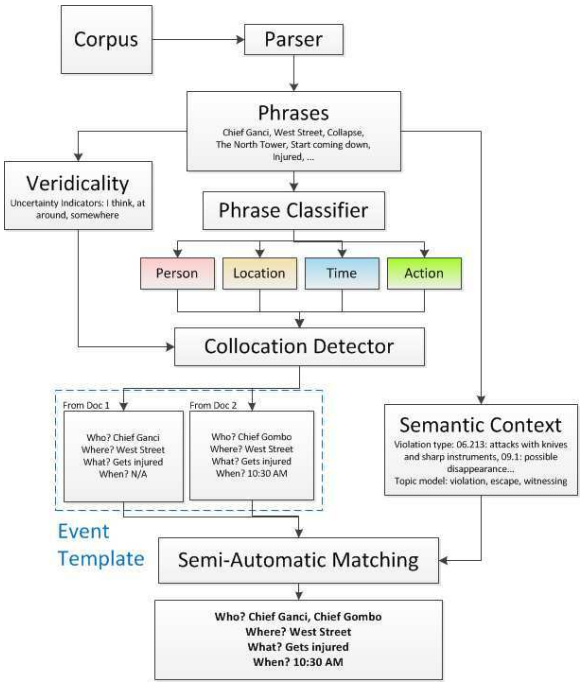

Figure 1. Model used in the NLP module for cross-referencing the entitites.

on from the extraction module, a phrase classifier is used to determine which phrases fall into important entity categories such as Person Names/Geographic Locations/Date/Time or depiction of an event. After classifying these phrases into categories, a module called Collocation Detector detects which of the entities are described in the same context within a passage in the corpus. This collocation of phrases is different from the collocation of words usually used in NLP in that it captures instances of the collocations, instead of a global probability. A collocation is only true when multiple elements correlate. After a set of collocated phrases have been detected, they are placed into the event template and fed into a visualization engine. Human observers then decide which cross-document entities are identical. The engine computes automatic scores to make suggestions to the observers on which events and entities should be merged.

2) Phrase extraction and classification: The first step of phrase extraction is done by running a full parser on each document and then extracting all the retrieved noun phrases and verb phrases from the parse tree. We decided against using a shallow parser (chunker) because it has lower recall (may not capture all the desired phrases) than a full parser. The parser we are using is the Stanford parser [31]. From the extracted phrases, we formulate a classification task for labeling important phrases for event extraction. The important phrase classification in our research is different from the traditional named entity recognition (NER) problem in NLP in that we are seeking to connect names to unnamed entities. We have 8 categories for important phrases: Organization, Person, Title, Location, Date, Time, Event, Miscellaneous, and the background category of Unimportant. Not all of these categories are visualized. Of these categories, some are traditional NER or TimeML categories. Event and Miscellaneous labels are new and determine some important phrases that might not be readily interpreted as named entities. Phrases such as "the pedestrian bridge," "the ferry," or "the second tower" which are not identifiable as a particular named entity but might be crucial in depicting the event are classified as Miscellaneous.

To maximally utilize human knowledge in the phrase labeling phase, an unsupervised selection mechanism selects the phrases to be labeled. In this mechanism, phrases are ranked by a score that is similar to a frequency or N-gram model, but it discounts the probability of a phrase if it is very common in a background corpus

$$\begin{aligned} S_c(phrase) = logP(phrase) \\ -max(logP_{bg}(phrase) \\ -logP(phrase), 0) \end{aligned} \quad (1)$$

where $logP(phrase)$ is computed by an N-gram language model trained on the current corpus, and $logP_{bg}(phrase)$ is based on a N-gram language model trained on a background corpus that is supposed to contain documents of all kinds. Under this model, the probability of a phrase is only discounted if $P_{bg}(phrase) > P(phrase)$. This application of Term Frequency Inverse Document Frequency[32] helps us to find frequent phrases in the corpus which are not popular in the background corpus. The phrases with top scores are manually labeled. Through this approach, we can obtain the labels for the most frequent and unique phrases in the corpus, which are likely to be more important in isolating an event. Our N-gram training uses the modified Kneser-Ney smoothing [33] from the MitLMpackage [34]. The background language model is obtained from Microsoft Web N-gram Services. Given a set of human-labeled phrases, we then train two levels of classifiers on these phrases. At the first level, a binary Important versus Unimportant phrase classifier is trained. At the second level, a one-against-all multi-class classifier is trained for each of the phrase categories described above, except Miscellaneous, which serves as the background category for important phrases. The features used for the classifiers are common NER features [35], [36], plus standard bag-of-words features. For the Date and Time phrases, we make use of the SUTime library [37] which matches date and time expressions using an extensive set of rules defined by regular expressions. The classification of these phrases do not depend on the labels.

For the collocation we use a simple metric: a Gaussian kernel on the distance between mentions of different phrases. Formally, the collocation probability of one occurrence of a phrase, given a set of other phrases is defined as

$$P(p_1|p_2,...,p_k) = exp(-\beta \sum_i (S(p_1) - S(p_2))^2) \quad (2)$$

where $S(p_i)$ is the sentence number where pi occurred. Given the defined conditionals, one can compute the joint probability $P(p1, p2, p3, ..., pk)$ and use a threshold to determine which phrase set goes to an event template.

IV. Presentation layer

The presentation layer consists of the visualization module and the feedback module. The visualization module consists of Storygraphs [38] and Storygrams. Storygraph, our earlier work, is a 2D visualization technique for presenting time and location on the same chart. It consists of two parallel vertical axes, which are used for latitude and longitude, and an orthogonal horizontal axis, which is used for time. An event, $E(lat, lng, time)$ is mapped in into the Storygraph by first drawing a line segment connecting the corresponding latitude and the longitude in two vertical axes. A marker is then placed on the line above the corresponding time of the event. Hence a pre-requisite for using Storygraph is that the data needs to be structured and precise, i.e it needs to have a precise geo-coordinates and timestamp. Any additional attributes like the type of event can be shown by changing the size, shape, and color of the marker.

One of the visualization layer goals is to present certain as well as uncertain data. Depending upon the context, uncertainty refers to semantic uncertainty, ranged values or missing data [39]. In our framework, uncertainty is introduced starting from the extraction phase as described above. These uncertainties include the temporal ("By this time, it had to be 11:00 o'clock at night" and "at that time I noticed"), locative, ("I guess that would be North End Avenue"), and entity ("At this point I had my five guys") [40]. We use Storygrams to present uncertainty.

A Storygram is a 2D-planar diagram consisting of events as its building blocks. An event in this case is a 3-tuple consisting of time, location, and entity. Storygrams are represented visually as triangles using the entities as vertices connected with weighted edges. The weights represent the confidence value in the relationship obtained from the NLP module. These confidence values show the certainty of the connection between two elements. Figure 2 shows a trigram with confidence values and elements.

To reduce visual clutter due to the excess of edges and vertices, we employ details on demand [41] and filtering capabilities on the dataset. The users can also drag and drop the events over other events and manually enter the confidence values to increase the association between two or more events. Since the users load one corpus at a time, this feature enables users to visually associate events in different documents.

The uncertainty in data also affects the output of semantic context and the verdicality modules. To address this issue,

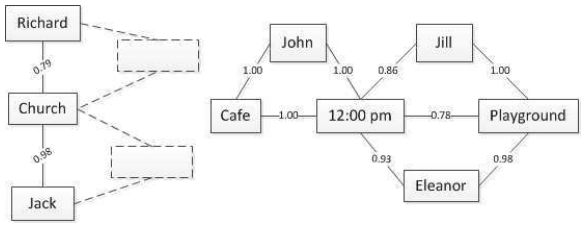

Figure 2. A model of Storygram showing events and uncertainty within the entities of the event.

we intend to make Storygrams interactive in that users can merge two vertices if they think they are the same. Merging of two vertices also marks the underlying data to be changed. The change is witheld until another expert verifies it. When the vertices are merged, the edges are also merged in many cases. In these cases the confidence values are also merged. This module is still under construction, and we leave it as a future work.

V. Discussion

Applying our framework to various datasets describing genocide in Cambodia from 1974-1979, a contemporary militant movement in central Africa, the attacks on the World Trade Center (WTC) in 2001, and U.S. military logs from Afghanistan has yielded visualization showing the movements of individuals and groups across the time and space of event contexts. Not per se indicative of violations, two of our corpora are included in this research because they are publicly available and share syntactic and semantic features to primary rights violations data. Those corpora are the WTC Task Force Interviews conducted with first responders to the attacks of September 11, 2001, and the material published by *The Guardian UK* as the "Afghanistan War Logs." At 511 interviews and 1.6m words, the first dataset does not qualify as "big," but the type of analysis our system affords is not replicable by non-computational methods as it relies on the cross-referencing of hundreds of thousands of granular phrases and the calculation of correlating uncertainties. Our analysis of the WTC material revealed that one individual, Father Mychal Judge, appeared 86 times in 33 interviews. These appearances describe his time on scene from when he arrived at the site to his death during the collapse of the north tower as bodies fell on the roof of 6 World Trade Center to his laying in state at St. Peter's. Extractions of these observations by survivors allow for the stitching together of a narrative describing the last day of Father Judge's life. We are still working on this corpus to extract narratives of other victims.

In the public reports from central Africa describing violations perpetrated by the Lord's Resistance Army [42], there are noticeable differences in reports regarding important details such as number and descriptions of victims and perpetrators, locations of incidences, and variations in time

and date. Similarly, detailed information may be given in ECCC interviews; however, key information such as what methods of interrogation and torture were permitted may vary depending on the individual and his or her position in the case. For example, discrepancies appear in the interview material in approved torture and interrogation techniques between Kaing Guek-Eav and Prak Khan[43], [44]. An important feature of our model is the conflation of multiple accounts describing a singular event. These multiple accounts provide a sense of depth and intricacy to the emerging broader narrative. Mining these details allows scholars to examine the multiple perspectives embedded in collective memory and cultural identity. Problems, however, abound when considering the scope of human rights violations as many narratives become part of the larger narrative and not always in the same form. From photos to memos to transcripts to field reports, portions of traumatic narratives live in a variety of documents that are not only multiplied by the number of atrocities, but also by the ways in which they are recounted. One such example can be found in the ECCC documentation and the euphemism used by the Khmer Rouge. In one of his interviews, Kaing Guek-Eav details the meanings of the key terms "smash" (execution), "resolve" (execution), "purge"(arrest), and "sweep cleanly away" [45]. Furthermore, his descriptions regarding which administrators used which terms expand the ability to reconcile other narratives to others. Through recognizing these terms and their variety of applications in regards to entities and times, one can better reconcile these voices with other narratives in history and help to shape identities within a given perspective.

As discussed in [38], our Storygraph visualization as seen in Figure 3 shows each individual report as a dot in a coordinate plane of latitude, longitude, and time. What our visualization revealed are both patterns in the corpus indicative of documentation practices and patterns in the event the reports describe. One such pattern, seen in the void spaces numbered 1, 2, and 3, corresponds to lulls in violence. Vertical banding as seen at A, B, and C, indicate events happening simultaneously across the geography described by the corpus. In this case, they indicated elections. In Figure 4, also discussed at length in [38], individual units within the corpus can be seen traversing the geography and the temporality of the corpus. Each colored line indicates one unit, and their path corresponds to the mentions within documents. An implication of these visualizations is that they flatten the corpus so as to enable a holistic perspective on the data, and may thereby lead researchers to focus on contextual analyses, rather than granular analysis of individual victims and perpetrators. However, a holistic view of human rights data sets such as we are undertaking with the confidential LRA material, when modelled on the approach seen in the Storygraph figures, facilitates drilling into the data. Seeing the pathway taken by an individual through a corpus when connected to the document fragments that were mined to produce that pathway deeply connects the documentary evidence to the final analytic visualization.

VI. Conclusion

This paper explored the challenges faced by researchers working on human rights violations, and proposed a framework for addressing some of those challenges. That framework included elements for processing the textual data, visualizing that data, and feeding judgments made by users of the framework back into the processing layer. Principally, our methods for information extraction, data visualization, and user feedback work to generate narratives that traverse document boundaries. These techniques for transversal reading of large historical corpora allow for the investigation of the relationship between big data and collective memories of traumatic events. Memory, as defined by Maurice Halbwachs in 1948 in the seminal work on collective memory [46], is by definition collective and only exists in the conversation undertaken by groups; individuals have no capacity for memory, only fantasy. This founding theory on collective memory corresponds to our projects argument related to cross-document coreference; information is only valid when it can be validated.

In the future, we plan to refine two areas of the framework and explore the potential applicability of this framework to corpora describing other categories of events. Areas of the framework that we plan to improve are our methods for determining the semantic context of Storygrams and our modeling of linguistic uncertainty and concomitant veridicality. Other categories of events that may be susceptible to this approach are environmental catastrophes, popular uprisings, and political movements.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. 1209172. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] V. Flusser, *Into the universe of technical images*. University of Minnesota Press, 2011, vol. 32.

[2] M. L. Best, W. J. Long, J. Etherton, and T. Smyth, "Rich digital media as a tool in post-conflict truth and reconciliation," *Media, War & Conflict*, vol. 4, no. 3, pp. 231–249, 2011.

[3] I. Norheim-Hagtun and P. Meier, "Crowdsourcing for crisis mapping in haiti," *innovations*, vol. 5, no. 4, pp. 81–89, 2010.

[4] R. Silva, J. Klingner, and S. Weikart, "State coordinated violence in chad under hissne habr: A report by benetech's human rights data analysis group to human rights watch and the chadian association of victims of political repression and crimes," Benetech, Tech. Rep., 2010.

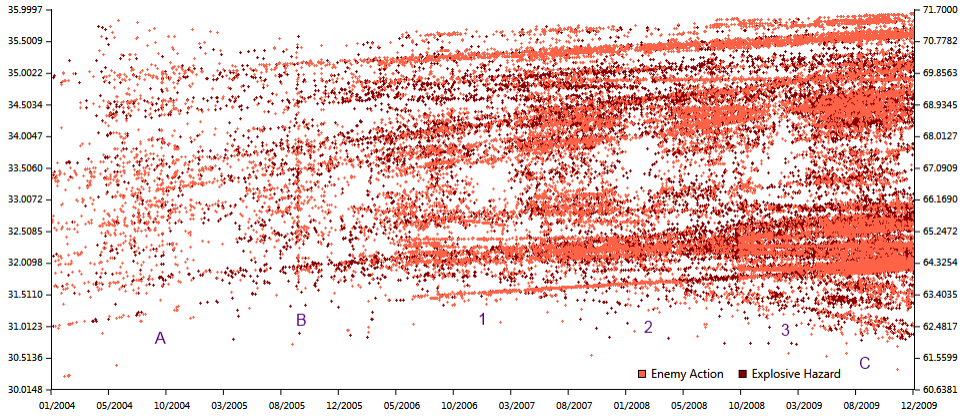

Figure 3. Figure adapted from our previous work [38] showing the movement of different units in the Afghanistan war.

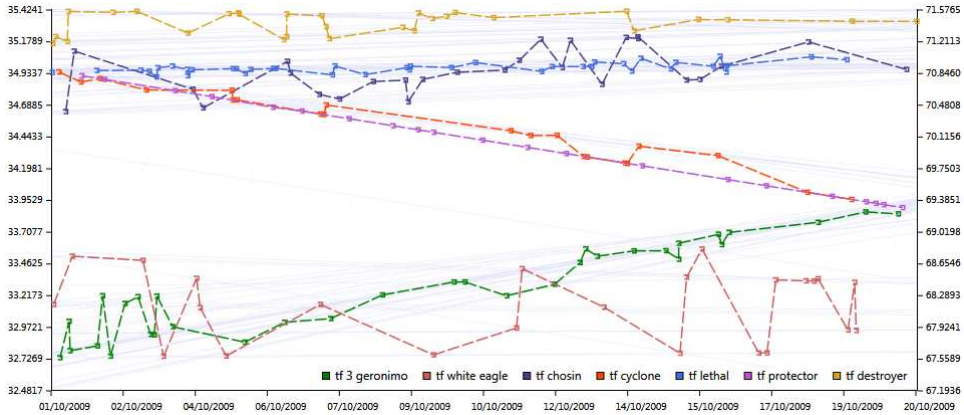

Figure 4. Figure adapted from our previous work [38] showing patterns in the Afghanistan war data. Patterns marked by $A - C$ show lots of events happening at the same time. $1 - 3$ show lack of events and periodicity of pauses.

[5] P. Ball, J. Asher, D. Sulmont, and D. Manrique, "How many peruvians have died?" American Association for the Advancement of Science, Tech. Rep., 2003.

[6] P. Ball, E. Tabeau, and P. Verwimp, "The bosnian book of dead: Assessment of the database (full report)," Households in Conflict Network, Tech. Rep., 2007.

[7] D. Boyd and K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," *Information, Communication & Society*, vol. 15, no. 5, pp. 662–679, 2012.

[8] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant *et al.*, "Quantitative analysis of culture using millions of digitized books," *science*, vol. 331, no. 6014, pp. 176–182, 2011.

[9] L. Manovich, "Data stream, database, timeline: the forms of social media," 2012. [Online]. Available: http://lab.softwarestudies.com/2012/10/data-stream-database-timeline-new.html

[10] Z. S. Syed, T. Finin, and A. Joshi, "Wikipedia as an ontology for describing documents." in *ICWSM*, 2008.

[11] M. L. Jockers, *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, 2013.

[12] S. Jones and H. Petrie, "Chartex: Discovering spatial descriptions and relationships in medieval charters," 2013.

[13] C. Grover, "Trading conferences," 2012.

[14] S. Soderland, B. Roof, B. Qin, S. Xu, O. Etzioni *et al.*, "Adapting open information extraction to domain-specific relations," *AI Magazine*, vol. 31, no. 3, pp. 93–102, 2010.

[15] Guardian.co.uk, "Afghanistan war logs," 2011. [Online]. Available: http://www.theguardian.com/world/the-war-logs

[16] M. Mateas and A. Stern, "Façade: An experiment in building a fully-realized interactive drama," in *Game Developers Conference, Game Design track*, vol. 2, 2003, p. 82.

[17] N. Kiyavitskaya, N. Zeni, J. R. Cordy, L. Mich, and J. Mylopoulos, "Cerno: Light-weight tool support for semantic annotation of textual documents," *Data & Knowledge Engineering*, vol. 68, no. 12, pp. 1470–1492, 2009.

[18] C. L. Sidner, "Towards a computational theory of definite anaphora comprehension in english discourse." DTIC Document, Tech. Rep., 1979.

[19] J. A. Hawkins, *Definiteness and Indefiniteness*. Humanities Press, 1978.

[20] J. Meyer and R. Dale, "Mining a corpus to support associative anaphora resolution," in *Proceedings of the Fourth International Conference on Discourse Anaphora and Anaphor Resolution*, 2002.

[21] R. Mitkov, B. Boguraev, and S. Lappin, "Introduction to the special issue on computational anaphora resolution," *Computational Linguistics*, vol. 27, no. 4, pp. 473–477, 2001.

[22] B. Webber, M. Egg, and V. Kordoni, "Discourse structure and language technology," *Natural Language Engineering*, vol. 18, no. 4, pp. 437–490, 2012.

[23] M. Dimitrov, K. Bontcheva, H. Cunningham, and D. Maynard, "A lightweight approach to coreference resolution for named entities in text," *Anaphora Processing: Linguistic, cognitive and computational modelling*, vol. 263, p. 97, 2005.

[24] S. Lappin and H. J. Leass, "An algorithm for pronominal anaphora resolution," *Computational linguistics*, vol. 20, no. 4, pp. 535–561, 1994.

[25] R. Mitkov, "Factors in anaphora resolution: they are not the only things that matter: a case study based on two different approaches," in *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*. Association for Computational Linguistics, 1997, pp. 14–21.

[26] K. Van Deemter and R. Kibble, "On coreferring: Coreference in muc and related annotation schemes," *Computational linguistics*, vol. 26, no. 4, pp. 629–637, 2000.

[27] C. Northwood, "Ternip: temporal expression recognition and normalisation in python," Ph.D. dissertation, Masters thesis, University of Sheffield, 2010.

[28] A. Auger and J. Roy, "Expression of uncertainty in linguistic data," in *11th International Conference on Information Fusion, 2008*. IEEE, 2008, pp. 1–8.

[29] M. J. Druzdzel, "Verbal uncertainty expressions: Literature review," *Pittsburgh, PA: Carnegie Mellon University, Department of Engineering and Public Policy*, 1989.

[30] E. Marshman, "Expressions of uncertainty in candidate knowledge-rich contexts: A comparison in english and french specialized texts," *Terminology*, vol. 14, no. 1, pp. 124–151, 2008.

[31] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 423–430.

[32] W. W. Cohen and J. Richman, "Learning to match and cluster large high-dimensional data sets for data integration," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 475–480.

[33] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proceedings of the 34th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1996, pp. 310–318.

[34] B.-J. Hsu and J. Glass, "Iterative language model estimation: efficient data structure & algorithms," in *Proceedings of Interspeech*, vol. 8, 2008, pp. 1–4.

[35] T. Zhang and D. Johnson, "A robust risk minimization based named entity recognition system," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. Association for Computational Linguistics, 2003, pp. 204–207.

[36] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2009, pp. 147–155.

[37] A. X. Chang and C. Manning, "Sutime: A library for recognizing and normalizing time expressions." in *LREC*, 2012, pp. 3735–3740.

[38] A. Shrestha, Y. Zhu, B. Miller, and Y. Zhao, "Storygraph: Telling stories from spatio-temporal data," in *Lecture Notes in Computer Science*, vol. 8034. Springer, 2013, pp. 693–703.

[39] A. T. Pang, C. M. Wittenbrink, and S. K. Lodha, "Approaches to uncertainty visualization," *The Visual Computer*, vol. 13, no. 8, pp. 370–390, 1997.

[40] W. T. C. T. Force, "World trade center task force interviews," 2001.

[41] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Visual Languages, 1996. Proceedings., IEEE Symposium on*. IEEE, 1996, pp. 336–343.

[42] C. Tracker, "Incidents export," Invisible Children and Resolve, Tech. Rep., 2013.

[43] ECCC, "Annex 25: Written record of interview - 09 june 1999," 1999.

[44] ECCC, "Written record of interview of Prak Khan," 2007.

[45] ECCC, "Written record of interview of Duch by CIJ on 21-01-2008," 2008.

[46] M. Halbwachs and L. A. Coser, *On collective memory*. University of Chicago Press, 1992.