Improving the Utility of Open-Source Event Data for the Design of Training Exercises

Doug Murphy, David Boyd, Dr. Bill Mandrick
InCadence Strategic Solutions
Fort Eustis, Virginia
dmurphy@incadencecorp.com
dboyd@ incadencecorp.com
wmandrick@incadencecorp.com

Duane "Dewey" Dannewitz
CGI Federal
Fort Eustis, Virginia
duane.dannewitz@cgifederal.com

ABSTRACT

The design of training exercises within the U.S Army involves simulating the real-world operational environment (OE) to the greatest extent possible. The increasing availability of public information relevant to the OE represents both opportunities and challenges for training and exercise planners. In this paper we examine using open-source event data as intelligence content by providing a faceted search capability based on an OE ontology. Event data is useful for exercise design not only because it informs the OE, but because events are used to motivate training tasks when the exercise is actually conducted. The news articles used as sources for event-mining systems can contain a variety of useful references to OE elements, including enumerations of intentional acts (violence, protests, diplomacy), and actors (victims, perpetrators) in a given area, as well as intangible variables such as effects upon regional stability and the mood of the populace. Many of these parameters are parsed and coded by event mining systems such as the Global Database of Events, Languages and Tone (GDELT) and the Integrated Crises Early Warning System (ICEWS). Further, these systems use a common taxonomy for event coding. Our OE ontology allows these data sets to be searched using terms that are familiar to unit commanders by mapping the event codes to OE taxonomies and lexicons used by the Army. We apply our ontology to build a search application hosted in our OE Enterprise Repository, a data store built on the Defense Information Systems Agency (DISA) Big Data Platform.

ABOUT THE AUTHORS

Doug Murphy is a Senior Software Engineer at InCadence Strategic Solutions Inc. currently developing enterprise web applications for the US Army. His previous experience includes work in systems engineering and data visualization at NASA, and creating interactive training tools for the US Navy. Mr. Murphy holds an MS in Electrical Engineering from the University of Southern California and a BS in Aerospace Engineering from the University of Michigan.

David Boyd is Vice President of Data Solutions at InCadence and has over 35 years of experience developing, integrating , and deploying complex data intensive systems in support of the DoD and IC communities. Mr. Boyd is a leading expert in Big Data systems and architectures, is the chair for the ANSI/INCITS Big Data technical committee and is a contributor and editor for several parts of ISO/IEC 20547 Information Technology – Big Data: Reference Architecture standard. Mr. Boyd is also a contributing author and editor for NIST Special Publications 1500-6 Big Data Interoperability Framework: Reference Architecture and 1500-7 Big Data Interoperability Framework: Standards Roadmap. Mr. Boyd holds an MS in Information Technology from Capella University and a BS in Computer Science from the University of Maryland, University College.

Dr. Bill Mandrick is a Senior Ontologist at InCadence Strategic Solutions and a Senior Research Fellow at the Joint Special Operations University (JSOU). His current research is on the development and use of standardized frameworks (ontologies) for organizing economic intelligence, specifically in the areas of land tenure and property rights (LTPR). He is also a Colonel in the U.S. Army Reserves with deployments to Iraq and Afghanistan where he has commanded soldiers, planned for major operations, and served as the primary civil-military operations advisor to a Brigade Combat Team.

Duane "Dewey" Dannewitz is a Director of Consulting and Project Manager for CGI Federal, Inc. currently supporting the US Army at the TRADOC G27 Operational Environment Training Support Center, Fort Eustis, Virginia. His previous experiences include serving as a career Counterintelligence and All-Source Intelligence Officer, TRADOC Instructor, Training Developer, Combat Developer, and Training Accreditor. Mr. Dannewitz holds an MS in Management from the University of Maryland, University College, and a BS in Human Anatomy and Cell Physiology from Minnesota State University, and is a graduate of the US Army Command and General Staff College.

Improving the Utility of Open-Source Event Data for the Design of Training Exercises

Doug Murphy, David Boyd, Dr. Bill Mandrick
InCadence Strategic Solutions
Fort Eustis, Virginia
dmurphy@incadencecorp.com
dboyd@ incadencecorp.com
wmandrick@incadencecorp.com

Duane "Dewey" Dannewitz
CGI Federal
Fort Eustis, Virginia
duane.dannewitz@cgifederal.com

INTRODUCTION

Background

In the late 1960s researchers began examining the idea that publicly available news reports could be used as data for a quantitative approach to foreign policy analysis (Schrodt, 1995). In recent decades, the growth of web accessible news combined with advances in automated systems for natural language processing have been a boon to this area of research, and has led to the success of several high-profile event mining projects (Beieler, 2016). Novel uses of event data, enabled in part by the greater accessibility of big data technologies to software developers and analysts, continue to appear. Our team, which supports the US Army TRADOC G27 Operational Environment Training Support Center (OE TSC) at Fort Eustis, is interested in using event data in the development of training exercises and simulations.

An area where quantitative foreign policy analysis overlaps with Army training is the need to understand the defining characteristics of a given region, both physical and intangible. Examples include the capabilities of the military, the demographics and strongly held beliefs of the population, and the stability of the government. From the perspective of a training developer, any such characteristic that bears on decisions of a military commander are part of the OE. Replicating the OE in the training environment is a key part of the home station exercise design process and factors heavily in the design of algorithms used in simulations. Because of this overlap between the design of event mining systems and the needs of Army training developers, event data contains features that can inform the OE. In this paper, we look at how this event data could serve training and simulation purposes more effectively.

Exercise/Training Design and the Operational Environment

The definition of the operational environment referenced by Army Training Circular 7-101 *Exercise Design* is "a composite of the conditions, circumstances, and influences that affect the employment of capabilities and bear on the decisions of the commander." To help the process of working inside such a broad subject, the OE is broken down into eight categories known as the operational variables. The operational variables are political, military, economic, social, information, infrastructure, physical environment, and time (known as PMESII-PT) (ADP 3-0, 2016). The eight variables and their sub variables describe the OE in terms that relate to specific situations as well as threat capabilities (Table 1).

The PMESII-PT variables produce a coherent profile of the exercise conditions that can be applied to multi-level training exercises and small-unit (brigade and below) training exercises alike. The Army's increased capability to integrate virtual, gaming, and constructive simulations with live training, along with the increased need for joint and multi-echelon training, requires exercise planners to develop a detailed and consistent OE for each training event. The process frequently includes creating storylines that are used to motivate the activities of the soldiers and role players when an exercise is conducted. Key events within such storylines may include riots, mass casualties, and missing personnel. TC 7-101 again advises the use of PMESII-PT to guide the development of these events. The G27 has recently completed a survey of simulation capabilities that leverage the PMESII-PT perspective, or something like it, in their implementation (Lee and Moyers, 2016). There are 23 active projects on this list.

Table 1. PMESII-PT variables and sub variables (TC 7-101, 2010)

Political Variable
- Attitude toward the United States
- Centers of political power
- Type of government
- Government effectiveness and legitimacy
- Influential political groups

Military Variable
- Military forces
- Government paramilitary forces
- Non-state paramilitary forces
- Unarmed combatants
- Nonmilitary armed combatants
- Military functions

Economic Variable
- Economic diversity
- Employment status
- Economic activity
- Illegal economic activity
- Banking and finance

Social Variable
- Demographic mix
- Social volatility
- Education level
- Ethnic diversity
- Religious diversity
- Population movement
- Common languages
- Criminal activity
- Human rights

Information Variable
- Public communications media
- Information warfare
- Intelligence
- Information management

Infrastructure Variable
- Construction pattern
- Urban zones
- Urbanized building density
- Utilities present
- Utility level
- Transportation architecture

Physical Environment Variable
- Terrain
- Observation and fields of fire
- Avenues of approach
- Key terrain
- Obstacles
- Cover and concealment
- Landforms
- Vegetation
- Terrain complexity
- Mobility classification
- Natural hazards
- Climate
- Weather
- Precipitation
- High temperature—heat index
- Low temperature—wind chill index
- Wind
- Visibility
- Cloud cover
- Relative humidity

Time Variable
- Knowledge of the AO
- Cultural perception of time
- Key event resolution
- Information offset
- Tactical exploitation of time
- Key dates, time periods, or events

In addition to PMESII-PT, there are several similar structures used by the Army for understanding the OE. The ASCOPE variables (areas, structures, capabilities, organizations, people and events) are used in civil affairs operations (FM 3-24, 2006). The recently developed Socio-Cultural Analysis Framework (SCAF) also focuses on the human domain (GCKN, 2016). The DIMEFIL (diplomatic, informational, military, economic, financial, intelligence and law enforcement) framework was devised to categorize "instruments of national power" (FM 3-05.130, 2008). Clearly the task of understanding the OE benefits from having a taxonomy for categorizing the topics of concern. For this reason we are interested in building a crosswalk between these taxonomies and the data produced by event mining systems.

Coded Event Data

An event, in the context of an event mining system, is a collection of coded fields distilled from a news article. Each coded event usually has three main parts: the source actor, the target actor, and the event type. An article describing a demand by the government of Iran for Saudi Arabian authorities to release an Iranian citizen from prison might be coded with the Iraqi prime-minister as the source actor, the Saudi Arabian envoy as the target actor, and use event type “demand release of persons or property.” Large repositories of such event data are available from a number of sources. An example developed by the US DoD is the Integrated Crisis Early Warning System (ICEWS), designed to provide situational understanding, forecasting and sentiment analysis (O’Brien, 2010). The Phoenix Project by the Open Event Data Alliance publishes both event data and the source code for the mining system itself (Schrodt et al., 2014). The Global Database of Events, Languages and Tone (GDELT), created by researcher Kalev Leetaru, publishes coded events and several other related data sets updated every quarter hour (Leetaru and Schrodt, 2013). While many other sources of event data exist, just these three contain tens of millions of events available for public use, including a subset of the ICEWS events.

Remarkably, the three projects listed above all use a common taxonomy for assigning event codes: the Conflict and Mediation Event Observations (CAMEO) framework. Released initially in 2002, CAMEO prescribes actor and verb codes in a who-did-what-to-whom format. The version of CAMEO used by GDELT contains just over 300 verb codes, ranging from the benign "Make Public Statement" (code 01) to the severe "Detonate Nuclear Weapons" (code 2042) (Schrodt, 2012). CAMEO also specifies an extensive system for codifying actors, including information such as nationality, ethnicity, religion and role (e.g. military, government, opposition). Strings of three-letter codes are used to identify each known attribute of the actor. These three-letter strings are concatenated to create actor codes of arbitrary length. For example, a sentence mentioning Syrian rebels could have an actor coding of SYRREB. Other common event fields include the location of the event, and metrics related to the severity of the action and the tone of the text in the source article.

The fact that CAMEO is so widely used benefits our effort to reuse event data for replicating the OE. Event verbs are clearly relevant to exercise storylines. The key event examples given in TC 7-101 can all be found nearly word-for-word in the CAMEO taxonomy. The CAMEO actor fields also match many of the factors needed to describe appropriate roles for training exercises and simulations. CAMEO therefore provides a promising conduit for our crosswalk between OE taxonomies and coded events.

UNDERSTANDING EVENTS IN AN OPERATIONAL ENVIRONMENT CONTEXT

Using Operational Variables as Metrics

Event data sets are useful as input to a wide range of simulation and analysis systems because they contain discrete, quantifiable elements. The actor and verb codes, geographic location data, and other metrics can all be counted, plotted, or otherwise measured in ways that would be difficult or impossible using only the original news articles. In a search interface, these parameters suggest themselves as facet fields, like the categories on shopping web sites that let the user filter their search results by price range or brand name. To further leverage the faceted approach specifically for the OE, we examined measuring the relevance of each CAMEO verb code against each operational variable in the PMESII-PT framework. A CAMEO code of 'Appeal for economic cooperation' for example, would score a maximum relevance to the economic variable in PMESII-PT, and less relevant to the other variables.

For our prototype search application, we applied a relevance scale of zero to four, with zero indicating no relevance between the event code and the operational variable, and four indicating the highest relevance. Because PMESII-PT has eight variables, each CAMEO code would receive an eight-digit relevance vector. By assigning numerical scores in this way, we can build a faceted search interface that presents a rough profile of OE relevance for any collection of coded events. Filtering events based on their PMESII-PT score enables a user to tailor the result set to the main OE variables of interest. These results can be used, for example, to inform the development of storylines and key events for exercises that emphasize a given OE variable.

A similar CAMEO mapping process can be applied using ASCOPE, DIMEFIL or other systems instead of or in addition to PMESII-PT. Because many of the tools used for operational and intelligence analysis rely on one or more of these categorization schemes, the mapping process can improve the utility of the event information. We believe this technique can effectively tailor the view of the OE along a given axis of interest, within a given framework, to benefit a variety of tools, methods and processes.

Assigning Relevance Scores

The process of assigning a relevance score from a CAMEO verb code to an operational PMESII-PT variable as described above is part objective analysis and part subjective interpretation. Many of the verb codes have a relatively clear emphasis on one or two PMESII-PT or ASCOPE variables, whereas others are less easy to score without additional information. However, because our primary use case in this work is to guide the process of finding events, we expect that some mismatches between the relevance scoring approach and the mindset of an end user should not completely overshadow the benefits. Still, the process of assigning scores should be conducted against some ground rules, or at least a consistent perspective. For this exercise, we assign relevance from the perspective of the commander tasked with developing a training exercise with key events well aligned to the OE.

Built-In Metrics

Our proposed relevance mapping technique scores CAMEO verbs purely on how well the event fits into the category described by an operational variable. This approach would, for example, assign the maximum relevance to PMESII-PT military for both verb codes “Use conventional military force” and “Ease military blockade.” GDELT, ICEWS and Phoenix events all include additional metrics that can help us further distinguish these events based on their impact to the OE. GDELT rates each verb code on the Goldstein scale, a number from -10 to +10 that measures the “impact to the stability of the region.” (Leetaru & Schrodt, 2013) The previous examples score -10 and +9 respectively on this scale. Each event is also assigned an “average tone” ranging from -100 to +100, assigned based on the severity or consequences of the event. Tone values consider the actual content of the article that generated the event, so that unlike the Goldstein scale, events with the same code can have different tone values. A massive riot would share the same event code as a smaller riot, but would receive a more negative tone value. We can leverage these metrics as additional search facets or to adjust the relevance scores.

IMPLEMENTATION

The OE Ontology

To implement our prototype event search interface, we extend a big data software system called the OE Enterprise Repository currently in the early stages of development at the OE TSC. A key component of this system is the OE ontology (OEO) developed by the G-27 and used to organize the objects in the database. We use an ontology because it facilitates semantic search, and because experience has shown that storing OE data lends itself to using graph-based structures. Also, ontologies are often used in causal and predictive analysis, topics of keen interest to both military and civilian researchers.

The OEO is a domain ontology designed to accommodate every kind of entity relevant to the operational environment. Because of the breadth of this topic, we extend the OEO from the Basic Formal Ontology (BFO) (Smith et al., 2005) and the Information Artifact Ontologies (IAO) (Smith et al., 2013). The BFO and IAO are established high level ontologies used in a variety of fields that will allow the OEO to benefit from the existing semantics. The IAO for example has classes for "Measurement Information Content Entity" and "Code", which can serve as parent classes for our relevance metrics and the CAMEO codes, respectively. A section of the OE ontology, shown with lineage through the IAO and BFO, is shown in Figure 1.

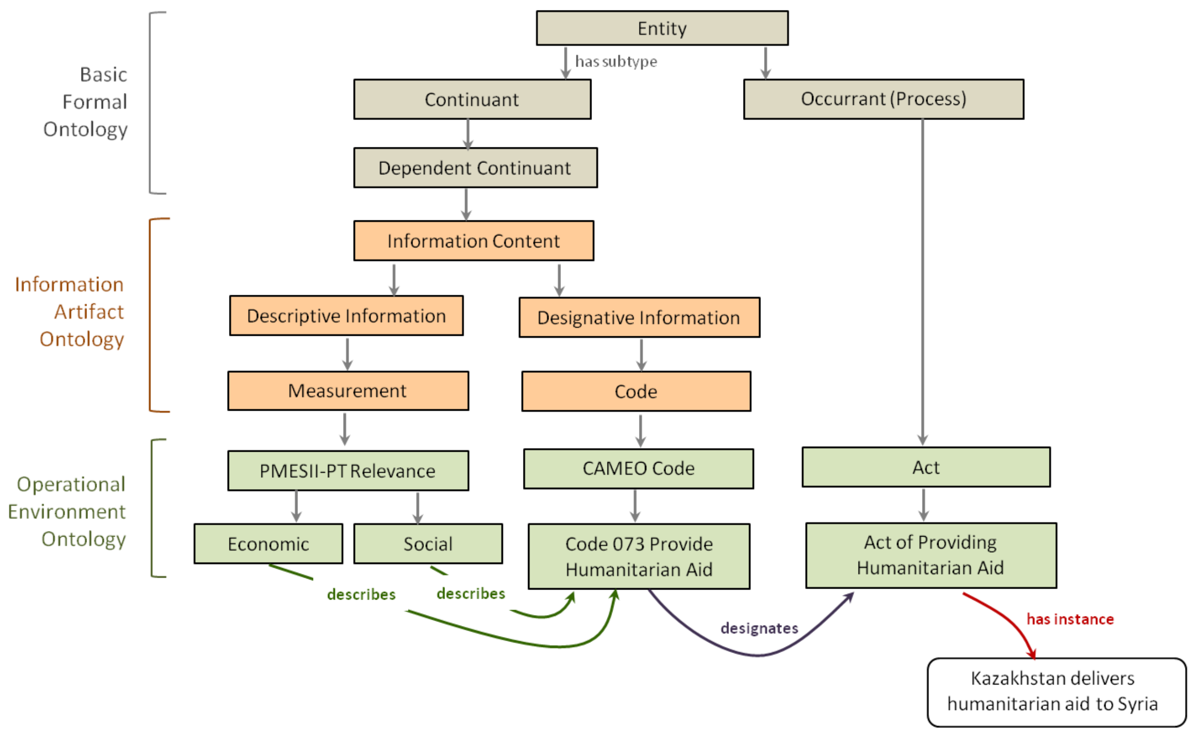

Figure 1. Rendering of ontology classes and relationships with one sample instance: "Kazakhstan delivers humanitarian aid to Syria."

With the proper placement of classes for events, event codes, and relevance codes in their respective ontologies, we then develop the relationships that will facilitate human or automated reasoning. This is done by adding relationships in addition to the standard hierarchical *has-subtype* relationship. For example, there are two important *is-about* relationships in the IAO: *designates* and *describes*. The *designates* relationship lets us capture the fact that a report about Kazakhstan delivering humanitarian aid to Syria would be assigned a CAMEO verb code "Act of Providing Humanitarian Aid." The *describes* relationship can capture any PMESII-PT score that has been assigned to this code. These connections also support faceted searching by activity type, event code, or relevance rating.

The OE Enterprise Repository and the Big Data Platform

To facilitate rapid ingestion and search of large data sets like the GDELT and ICEWS event data, the OE Enterprise Repository required a horizontally scalable big data architecture. We selected the DISA Big Data Platform (BDP) (Landreth & Bart, 2015) as our baseline framework. The BDP was originally developed by DISA to support cyber security analysis as part of the Joint Regional Security Stack (JRSS). Key characteristics of the BDP stack that led to its selection for our efforts include:

- Accredited for processing classified information
- Modular architecture that allows us to only install the required components and to add new components as required.
- A complete semantic data stack based on the Navy's RYA implementation of RDF on Accumulo (Crainiceanu et al., 2012, Punnoose et al., 2015).

On top of the BDP we are using an entity meta-model framework developed by InCadence called Coalesce. Coalesce provides a flexible approach to entity definitions and attributes through a consistent well defined application programming interface. Coalesce also has a modular data persistence architecture that allows multiple persistence mechanisms to be used in parallel to support optimized access to the data based on the application requirements.

Our long-term objective with this architecture is to deliver a fully linked knowledge graph representation of the OE that supports search, reasoning, and analytics. Figure 2 shows the key elements of the OE Enterprise Repository.

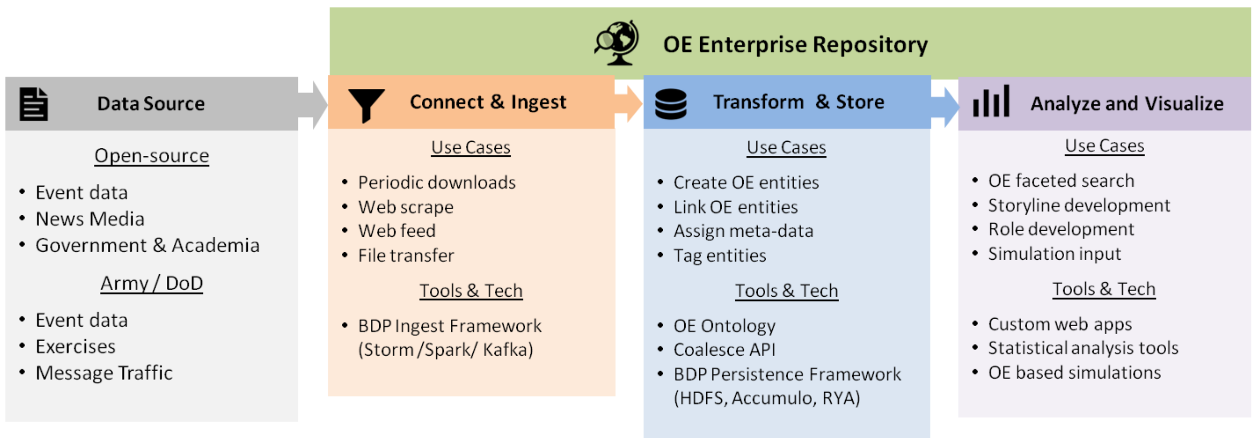

Figure 2. The OE Enterprise Repository architecture overview with sample data set types, important uses cases and key enabling tools/technologies identified.

A Prototype Search UI

To show the operational utility of using OE variables as facets, we developed a map-based search tool for coded events. The events were downloaded from the GDELT project website and ingested into the OE Enterprise Repository, and PMESII-PT relevance scores were assigned to each CAMEO code. The search interface is shown in Figure 3.

Circular event markers are rendered on the map at the coded event locations, colored based on positive or negative average tone (red/blue) or Goldstein scale score (not shown). Circle diameter is proportional to the magnitude of the average tone or Goldstein scale value. The bar chart in the upper left shows the average score of all the events currently drawn on the map in each of the PMESII-PT variables. The sliders under the bar chart filter the event data, hiding any event whose PMESII-PT score is lower than the slider value. Once the filters have been applied, the bar chart is updated based on the scores of the remaining events. Individual event descriptions can be viewed by clicking the markers. Actor demographics such as ethnicity, religion and role are rendered as pie charts.

With this interface we are able, through the process of filtering and rendering, to quickly tune the result set to match a desired OE profile. Even a few iterations can narrow tens of thousands of events into a small collection that could easily be read through by the user. For recent events GDELT often provides the URL of the original news item. For some use cases, such as authoring key events in an exercise, we expect the ability to find and examine the original news articles could be the end goal. In this case PMESII-PT filtering simply provides a potential shortcut to web based news search. In other cases, where collections of events or actors are needed as inputs to a simulation, the coded data itself or the meta-data could prove useful.

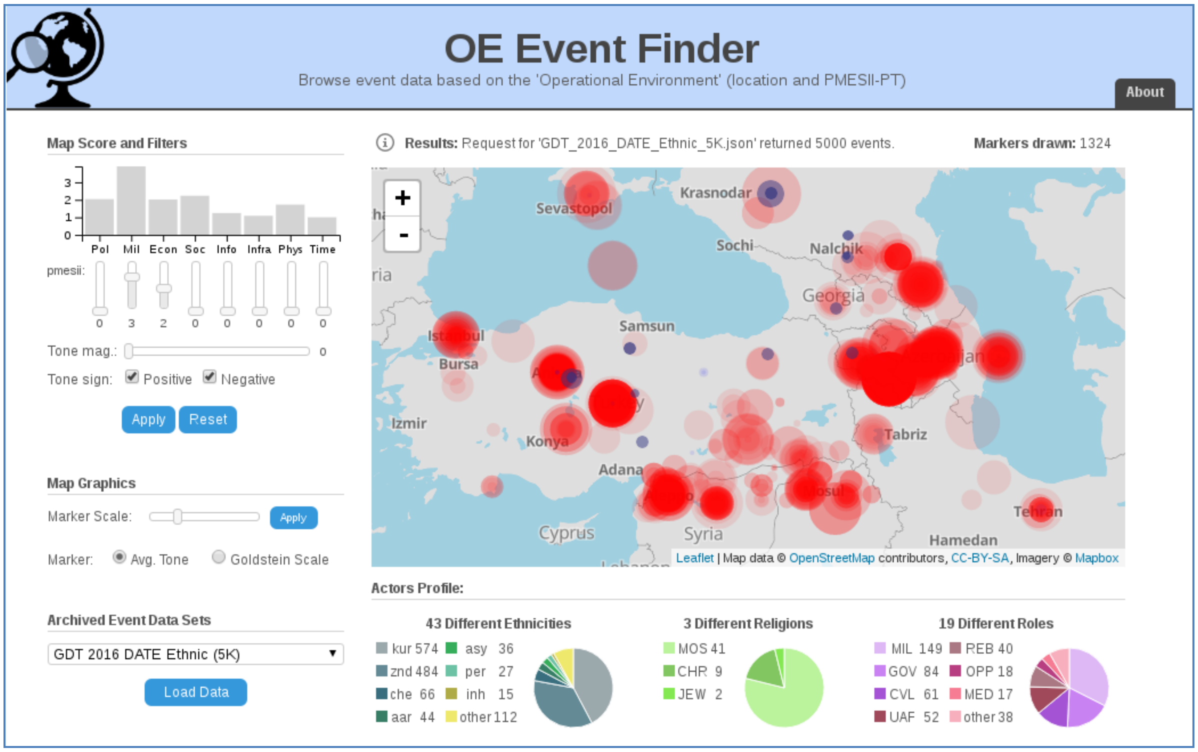

Figure 3: Elements of an Event Search Interface Using PMESII-PT Facets

Another effect of browsing event data and reading the original source news is that the user gains some insight into the quality of the coding systems. Coding accuracy rates are sometimes documented by the mining system developers or other researchers (Schrodt, 2015). These rates should be considered when using event data for any process without a human in the loop.

CONCLUSIONS AND FUTURE WORK

Looking at coded event data through the lens of operational variables such as PMESII-PT has several advantages over traditional web search for the purposes of building training exercises and simulations. It allows a faceted filtering of a large data set along an axis of interest to the end user, even though the names of the variables may never appear in the event description. The combination of map features and tone metrics also provides capabilities not readily available in a news-search approach. Shortcomings of the coded events such as miscoding and irrelevance can still impact the utility of this system. Techniques for minimizing this impact through interface features or additional algorithms will likely be informed by the needs of the specific use case.

In the survey of simulations and tools conducted by Lee and Moyers, frequent mention is made of event related concepts such as actors, actions, mood, and sentiment. Some of these tools already make use of or produce coded events. Using our mapping technique could benefit the ability of these systems to respond to users with an interest in a set of operational variables. Identifying or creating real or fictitious role players or organizations is part of the preparation of input data for some of these tools, in addition to being a key part of exercise design. The extensive actor data associated with coded events seems a promising source of information to populate this input. Further, several of the tools identified in the survey have a causal and predictive analysis component (ICEWS among them). Effective predictive analytics and causal analysis will require a wide range of data extraction and tagging techniques. Coding of events against a semantically consistent ontology, and scoring events along vectors (e.g.

PMESII, ASCOPE) will provide additional information against which predictive analytics can operate. Codes and scores as we have described provide for natural clustering of events along well defined dimensions which will support analytics.

Our experience with the prototype interface raised several potentially interesting areas of future work. The OE mapping process might be improved by considering both the actor codes and the event codes to build a PMESII-PT vector. A code of "attempt to assassinate" could be more or less relevant to the military, political, social, and even economic variables depending on who made the attempt or the identity of the target. Also, the many of the OE categorization systems or loose taxonomies are hierarchical. PMESII-PT and SCAF for example both have sub-variables under each root category. Extending the CAMEO mapping down to these sub variables, combined with using actor codes to inform the relevance score, could provide greater resolution for specific use cases. Long term analytic approaches that leverage actor codes, sentiment and other variables can be defined to automatically compute the relevance scores. In addition, by storing the actors and events in a linked semantic store, analytics can be developed to explore the causal impacts of actors and events on the OE.

ACKNOWLEDGEMENTS

The authors would like to thank Mr. Tony Cerri, Director of the Data Science, Models, Simulations Lab at the TRADOC G27 OE TSC, for his support during this effort, and for fostering an environment that encourages creativity.

REFERENCES

Beieler, J. (2016). Creating a Real-Time, Reproducible Event Dataset. *arXiv preprint arXiv:1612.00866*.

Crainiceanu, A., Punnoose, R., & Rapp, D. (2012). Rya: A Scalable RDF Triple Store For The Clouds. in *Cloud-I Workshop*.

HQ Department of the Army. (2016, November). *Army Doctrine Publication (ADP) 3-0 Operations.*

HQ Department of the Army. (2008, September). *Field Manual (FM) 3-05.130 Army Special Operations Forces Unconventional Warfare.*

HQ Department of the Army. (2014, May).). *Field Manual (FM) 6-0 Commander and Staff Organization and Operations.*

HQ Department of the Army. (2006, December). *Field Manual (FM) 3-24 Counterinsurgency*.

HQ Department of the Army. (2010, November 26). *Training Circular (TC) 7-101 Exercise Design*.

Landreth, B., & Bart, D. (2015). Cyber Situational Awareness - Big Data Solution. Retrieved December 1, 2016 from http://www.disa.mil/~/media/Files/DISA/News/Conference/2015/Defend_CyberSABigDataSolutionLandreth.pdf

Lee, H., & Moyers, C. (2016). *Compendium of PMESII Analysis Methods, Models and Tools.* TRADOC G-27 OE Training Support Center, Fort Leavenworth, KS.

Leetaru, K., & Schrodt, P. A. (2013, April). GDELT: Global Data on Events, Location, and Tone, 1979–2012. In *ISA Annual Convention* (Vol. 2, No. 4).

O'Brien, S. P. (2010), Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research. *International Studies Review*, 12: 87–104.

Punnoose, R., Crainiceanu, A., & Rapp, D. (2015). SPARQL in the cloud using Rya. *Information Systems, 48*, 181-195.

Schrodt, P. A. (1995). Event data in foreign policy analysis. *Foreign Policy Analysis: Continuity and Change in Its Second Generation*, 145-166.

Schrodt, P. A. (2012). CAMEO Conflict and Mediation Event Observations Event and Actor Codebook Version 1.1b3. Retrieved November 11, 2016 from http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf

Schrodt, P.A., Beieler, J., & Idris, M. (2014, March) Three's a Charm?: Open Event Data Coding with EL:DIABLO, PETRARCH, and the Open Event Data Alliance. Paper presented at the International Studies Association meetings, Toronto.

Schrodt, P. A. (2015, June 25). Comparison Metrics for Large Scale Political Event Data Sets. Paper presented at the 2015 European Political Science Association meetings, Vienna.

Smith, B., Kumar, A., & Bittner, T. (2005). Basic formal ontology for Bioinformatics. *Journal of Information Systems*, 1-16.

Smith, B., Malyuta, T., Rudnicki, R., Mandrick, W., Salmen, D., Morosoff, P., ... & Parent, K. (2013). IAO-Intel An Ontology of Information Artifacts in the Intelligence Domain.

TRADOC G-20 Global Cultural Knowledge Network (GCKN). (2016, October). *Socio-Cultural Analysis Framework (SCAF)*. Ft. Leavenworth, KS.