A New, Near-Real-Time Event Dataset and the Role of Versioning [Draft]

Andrew Halterman
Caerus Associates
ahalterman0@gmail.com

John Beieler
Pennsylvania State University
johnb30@gmail.com

We describe a new event dataset, Phoenix, and discuss the issues we faced in the process of generating and managing the dataset. Phoenix records dyadic political and social events coded with the CAMEO scheme with global daily coverage from the 1970s to the present. The core objectives for the dataset include accuracy, transparency, and replicability. To this end, we have created a number of new programs to generate the data, which we have released under an open source license for others to use. Part of our hope in releasing all of the tools and pipelines to generate the data is others will create more custom-built datasets tailored to specific subjects and to bring a broader community into to the creation (not just use) of event data. One challenge of rapid, community-driven updates to the tools and datasets is that they can cause confusion and churn in the data. We argue that two communities of users exist, with different desires for stability and frequent updates, and that their needs can best be served by throughly versioning the data and the software and dictionaries used to create it.

1 Introducing Phoenix

The two of us, along with Phil Schrodt, Patrick Brandt, and Erin Simpson, have constructed a new, daily updated event data set with global coverage called Phoenix (Beieler et al. (forthcoming)). Phoenix, PETRARCH, and the associated other software are created and developed by the Open Event Data Alliance, a new research consortium for researchers engaged in the production, evaluation, and use of political event data.[1]

Event data records political and social interactions between actors in a way that allows for easy aggregation and statistical analysis. In almost all cases, event data consists of a source actor, an action, and a target actor. For instance, take this example sentence:

> "The United States launched an air strike against an Islamic State tar-

[1]The Open Event Data Alliance core membership consists of Phil Schrodt of Parus Analytics, Andrew Halterman and Erin Simpson of Caerus Associates, John Beieler of Pennslyvania State University, and Patrick Brandt of the University of Texas at Dallas. The OEDA website is located at `http://openeventdata.org/` and OEDA software and dictionaries are at `https://github.com/openeventdata/`. We gratefully acknowlege the major contributions of the other three members to the ideas we present in this paper.

> get southwest of Baghdad, the U.S. Central Command said on Monday night, expanding its campaign against the militant group that has seized parts of Iraq and Syria."[2]

An event data system would recognize the source actor as the United States, the event as some kind of use of aerial weapons, and the target actor as the Islamic State or Syrian rebels. In our dataset, for instance, the event is represented as

> USA 195 IMGMUSISI

(the latter actor code is for an international militarized group, with the attribute code for Muslim and a special group designator ISI that corresponds to ISIS/ISIL/the Islamic State).

The Phoenix dataset is coded using the PETRARCH coder, a Python-based program using full parsing of news text, replacing the earlier and less sophisticated TABARI software (P. Schrodt 2001). PETRARCH uses Stanford's CoreNLP to do deep parsing of news text and should be much more accurate and extendable than TABARI. The three big advantages of PETRARCH are in the deep parsing of of sentences into their parts of speech (as Phil Schrodt puts it, parsing should be done by computational linguists, not political scientists), the ability to easily add or swap out modules (such as geolocation, event disambiguation, issue tagging), and the much greater ease of reading and contributing to the code.

Currently, Phoenix includes approximately 250,000 events generated from hourly scrapes of 450 English language news sites over the past 100 days. We are currently in the process of parsing and coding a corpus of historical documents running back to the 1970s, which will be foundation of the initial release of the Phoenix dataset, planned for within the next month.

Like many other automatically coded event datasets, Phoenix is heavily skewed toward US coverage, though we hope to reduce this as much as possible by including international reporting and local news sources as much as possible.

Phoenix displays a similar distribution of CAMEO event types as other CAMEO datasets. Event type 01 ("Make public statement") and 04 ("Consult") are heavily represented, with the third largest category being 19 ("Fight"). Much of the recent popular attention toward event data has been directed at is coverage of protests (CAMEO code 14), which are a remarkably small fraction of Phoenix's coverage.

[2] Jason Szep and Mehrdad Balali, "Iran supreme leader spurns U.S. overture to fight Islamic State," *Reuters*, September 16, 2014, `http://www.reuters.com/article/2014/09/16/us-iraq-crisis-idUSKBN0HA1SD20140916`

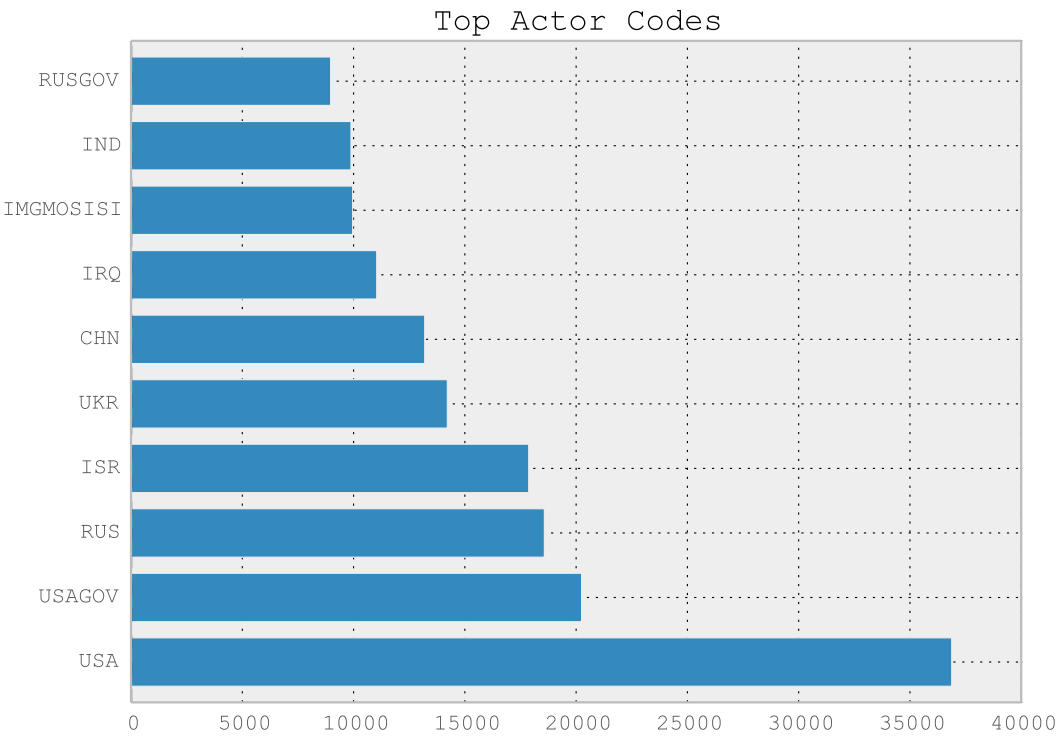

Figure 1: Count of Actor Codes in Phoenix

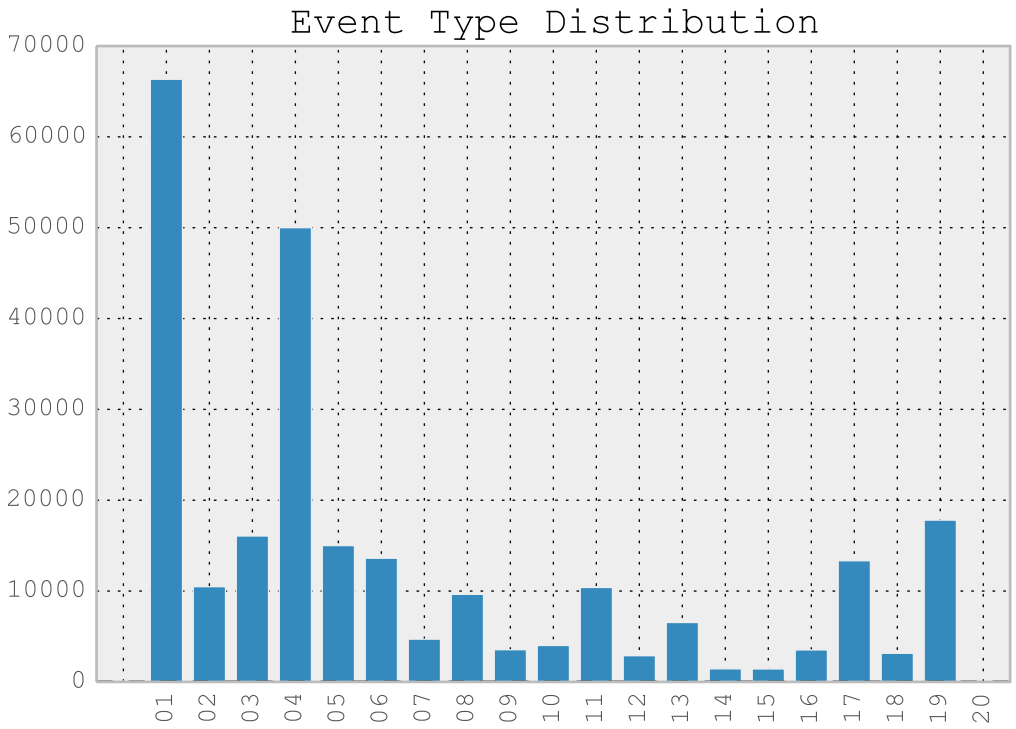

Figure 2: Distribution of CAMEO Event Codes in Phoenix

2 Opportunities and Challenges with Phoenix

Phoenix includes several features and objectives that have not been fully implemented in previous event datasets. At the same time, we have chosen to not implement some of the components that are present in other datasets, specifically GDELT. The most significant omission we have made is geolocation, though we hope to have it added as soon as we are satisfied with the accuracy of our system. GDELT, as the first automatically geolocated, publicly available machine-coded dataset, greatly changed the popular appeal of event data and led to several interesting projects using its geolocation data.[3] The future of automated event data undoubtedly includes geolocation, but we have chosen to omit it from the first version of Phoenix because the existing, publicly available geocoding software[4] was not accurate enough in determining where events described in text occured. We have experimented with Penn State GeoVista's GeoTxt software (Karimzadeh et al. 2013) and with Berico Technologies' CLAVIN and MIT's modification of CLAVIN, called CLIFF (D'Ignazio et al. 2014). However, varying degrees of inaccurate output or difficulty in setting up and running made each of them unfeasable for our use. A top research priority for our group is developing open-source software to geolocate events described in text.

A second component of GDELT that we have decided to not implement is a comprehensive ingestion of all English language news reporting present on Google News. Phoenix uses a whitelist of around 450 English language news sources, which are further categorized as wire services, international papers, and local news sources. Phoenix's software, like GDELT's, is designed to extract major international political and social events from wire service reports. Our exploration of GDELT found many instances of mis-codings from ambiguously-worded local US papers. We do not include all English language reporting, but we have also moved beyond the 1-4 wire services used in KEDs and other event data projects. Relatedly, while GDELT coded the entire text of each story, we code only the first several sentences (the exact number can be specified in the coding software) after finding that later sentences tend to include almost entirely historical background or news analysis that cannot be coding in an event data framework.

[3] It was also used for several analytically poor pieces that assumed that GDELT's geolocated representation of events was comprehensive and completely accurate. See for instance, FiveThirtyEight's retraction of a piece using GDELT's geolocation to quantify kidnapping in Nigeria: `http://fivethirtyeight.com/datalab/mapping-kidnappings-in-nigeria/`

[4] GDELT's geocoding software is not publicly available and we were not satisfied with its geolocation accuracy.

2.1 Transparency in the Data Generating Process

One top priority in creating Phoenix was to ensure complete transparency into the data generating process. In order to be useful to and trusted by researchers, the dataset needs to be traceable from input source text to the resulting event data. Small changes in the coding process can lead to large changes in the output, making the ability to inspect all of the pipeline components crucial. In our case, we have tried to achieve this by making all of our software and our dictionaries available online in an easily installable format, with permissive licenses that allow anyone to copy and distribute our software for almost any purpose, including commerical use.

One of our hopes in distributing our software and dictionaries is that other people will use them to generate their own event data or will contribute improvements to ours. As this happens, we will need to incorporate contributions and improvements in an orderly way, which will necessitate data management procedures like the ones we describe below. In turn, these data management systems depend on users being able to trace data back to the exact tools used to generate it, which relies on being able to freely access those tools. We hope that this interlocking system of technical transparency will create a virtuous cycle of improving accuracy, increasing use, and continuing development.

One of the promises of automated event data generation (as opposed to human generation) is that as dictionaries are improved, new source texts are acquired, and coding ontologies are sharpened or expanded, the coding software can be re-run over previously coded input texts to geneate updated event data. While this has happened internally and on small scales as part of the process of refining the codebook and dictionaries (P. A. Schrodt, Simpson, and Gerner 2001, 40), major machine-generated event datasets have not been updated over time. If the open source software model can be successfully applied to event data, we can harness the work of people building custom datasets for their own projects to improve the event data available to everyone. But continuous updates to datasets are not always desireable.

3 The Pitfalls Of Updating Data

Setting the techincal obstacles aside, one of the greatest challenges we faced was in decided how to manage the tradeoff between rapid innovation and consistency. Different users of our dataset have greatly differing demands.The first group consists of people using the data primarily for monitoring, reporting, and one-off, short term projects. The advent of free, daily updated event datasets has led to many government agencies, NGOs, and research organizations using the data for situtational monitoring. These users are generally best served by constant, immediate

improvements to the dictionaries, coding software, and number of sources as they occur. Many of the updates to our dictionaries have been driven by ongoing news, as we opportunistically add groups like ISIS and take out overly broad verbs that become apparent during sporting events like the World Cup and accidental airplane crashes like MH370 in the Indian Ocean. These targeted updates create day-to-day variation in our coverage, but for monitoring uses, the improved coverage is worth the variation.

Other users of event data, including almost all academic users, require stability in the data generating process to allow causal inference, accurate forecasts, and long-term research projects. These users require assurances that a particular version of the source list, dictionaries, and coding software will have created event data for the entire period of our coverage and will continue to do so into the future.

To demonstrate the large discontinuities that can result from changes in the coding process, we show differences in two types of events that were the subject of extensive dictionary updates during the summer: Syrian rebel groups, and the false positives for conventional military force (CAMEO 19) that were highlighted during the World Cup.[5]

One place where we expended effort on updating dictionaries was for rebel groups in Syria. Before the update, the dictionaries used in KEDS, the Penn State event data projects, and GDELT did not include any of the rebel groups fighting in Syria, meaning that news reports that used their names to describe events would not be coded. Unsurprisingly, adding the rebel groups to the dictionaries produced a marked increase in the number of events coded with Syrian rebels as the source actor.

A second component of the dictionaries that recieved a large amount of work this summer was the verb phrases for the "fight" events (CAMEO code 19), and specifically for 190 ("Use conventional military force, not specified 222 below"). Early user feedback indicated that many of the events that were coded as "fight" were descriptions of World Cup matches or metaphorical uses of verb phrases involving conflict and military action. Several dozen changes to the verb dictionary noticably changed the overall number of events.[6]

But merely looking at the overall number of "fight" events does not indicate whether we've successfully reduced false positives without sacrificing sensitivity. Short of

[5]All of the changes made to our dictionaries since the formation of the OEDA in the spring of 2014 are visible on our Github page: `https://github.com/openeventdata/Dictionaries`

[6]The changes are visible in these three commits to the repository on Github: `https://github.com/openeventdata/Dictionaries/commit/e0c0a380abd422cb96a90afe8e7473d42997a847`, `https://github.com/openeventdata/Dictionaries/commit/d7cf0d83ff6247c537fea3dcca6fdf92e59fdba5`, and `https://github.com/openeventdata/Dictionaries/commit/21c0bb259e1d841c3589824ec6dd035dedeef77e`

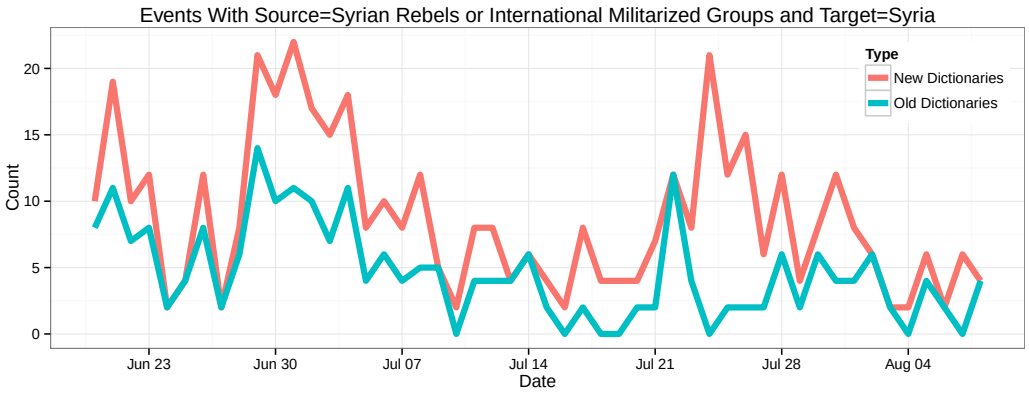

Figure 3: Expansion of the dictionaries for Syria Unsurprisingly Increases the Number of Events in Syria

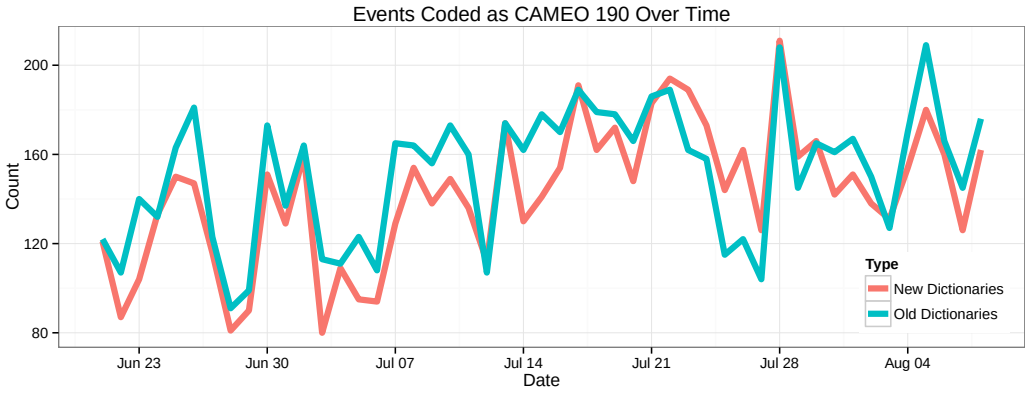

Figure 4: Reductions in CAMEO 190s ("Use conventional military force, not specified below") After Dictionary Updates

manually verifying the events, one way of estimating how useful the changes were is to see which countries saw their number of conflict events increase and decrease.

Table 1: Changes in Codings of CAMEO 190s (Fight; unspecified) Between Old and Updated Dictionaries.

Country	Old Count	New Count	Number Change	Percent Change
United States	738	651	-87	-11.8%
France	122	66	-56	-45.9
Germany	86	35	-51	-59.3
United Kingdom	149	116	-33	-22.1
Italy	50	18	-32	-64.0
Syrian Arab Republic	260	233	-27	-10.4
. . .	. . .	. . .	. . .	. . .
Intergovernmental Organization	95	122	+27	+28.4
International Militarized Group	94	156	+62	+66.0
Pakistan	244	307	+63	+25.8
Israel	1268	1345	+77	+6.1
Palestine, State of	672	757	+85	+12.6
Russian Federation	286	385	+99	+34.6

Table 1 shows that with the exception of Syria, all of the countries that saw their "fight" events decrease are developed, stable countries that are not experiencing conflict inside their borders (though their troops are engaged in foreign conflict in some cases, which would lead to events where they are coded as the source actor in a code 19). Likewise, with the exception of "Intergovernmental Organization," all of the actors that saw their number of "fight" events increase are engaged in ongoing conflict in their territory or a neighboring territory.

4 Versioning As A Solution to Data Churn

The solution we have settled on for the problem of frequent changes from improved dictionaries and software is to version, tag, and release all of our software, dictionaries, and source lists in a way that allows people to associate any entry in one of our datasets with the exact process used to generate it. We plan to make a large new release of the data every six months, and commit to producing that version of the data daily for the next year. We also plan to make a bleeding-edge version available that incorporates changes as soon as we make them so people wanting the most up-to-date features will have access to them. Thus, we expect to be maintaining and producing three datasets at any given time.

In its technical implementation, our versioning system roughly follows the semantic

versioning approach ("Semantic Versioning 2.0.0"), which uses version numbers that look like "1.0.0".

1. The first number is a major version change that majorly breaks backward compatability (represented by a shift from, e.g., 1.0.0 to 2.0.0). We would use this version number to indicate a major change in PETRARCH or other pieces of software used in the pipeline, the adoption of a coding scheme besides CAMEO, or when column names and numbers change in a way that would break people's code and databases.

2. The second number in the Semantic Versioning system is for minor updates, and is represented by an incrementing from 1.1.0 to 1.2.0). We would increase this version number any time we re-coded our entire corpus of text with new dictionaries, sources, and small modifications and made a committment to run it forward for the next year. These changes would not drastically change the structure of the data (column names and types would remain the same) but could break forecasting and inferential models that were trained on previous versions of the data.

3. "Patches", the final number (e.g. 2.0.1 to 2.0.2) reflect each incremental improvement as it's added to the code base. These version numbers would increase any time a contributer updated the dictionaries or code, and which would be reflected in the daily, cutting edge version of the dataset. These patch versions would not be applied to the historical corpus of text and would only be used until the next patch to the software. When the time came to release a new minor version, all of the patches that were added after the previous minor release would be folded together into the new minor release.

The role of transparency: Critical for our system of versioning to work is that people can view all of our software and coding materials and all of the chagnes we make between releases. We have made a strong committment to making our software and data free and open source in perpetuity, and this committment is what allows us to use a system of versioning.

5 Conclusion

We describe a new daily-updated event data set Phoenix, and some of the issues we have faced in creating it. One novel problem is the need to incorporate frequent changes made to the dictionaries and software by many contributers in a way that maximized people's ability to use the highest quality data available while ensuring enough data stability to permit causal inference and forecasting. We argue that this balance can best be met by releasing all changes as they are added to the users who would benefit from constant updates, and by rolling changes into larger

releases twice a year and applying those changes to our entire backfiles of news text. Versioning both the data and the tools used to create it allows anyone to associate an event in our dataset with the exact software and dictionary configuration that created it, which is important for ensuring validity and consistency.

Bibliography

Beieler, John, Patrick T. Brandt, Andrew Halterman, Philip A Schrodt, and Erin M. Simpson. (forthcoming). "Generating Political Event Data in Near Real Time: Opportunities and Challenges." In *Data Analytics in Social Science, Government, and Industry*, edited by R. Michael Alvarez. Cambridge University Press.

D'Ignazio, Catherine, Rahul Bhargava, Ethan Zuckerman, and Luisa Beck. 2014. "CLIFF-CLAVIN: Determining Geographic Focus for News." In *"KDD 2014: New York, New York, August 24, 2014"*.

Karimzadeh, Morteza, Wenyi Huang, Siddhartha Banerjee, Jan Oliver Wallgrn, Frank Hardisty, Scott Pezanowski, Prasenjit Mitra, and Alan M MacEachren. 2013. "GeoTxt: a Web API to Leverage Place References in Text." In *Proceedings of the 7th Workshop on Geographic Information Retrieval*, 72–73. ACM.

Schrodt, Philip. 2001. "Automated Coding of International Event Data Using Sparse Parsing Techniques." In *annual Meeting of the International Studies Association, Chicago*.

Schrodt, Philip A., Erin M. Simpson, and Deborah J. Gerner. 2001. "Monitoring Conflict Using Automated Coding of Newswire Reports: a Comparison of Five Geographical Regions." In *PRIO/Uppsala University/DECRG High-Level Scientific Conference on Identifying Wars: Systematic Conflict Research and Its Utility in Conflict Resolution and Prevention, Uppsala, Sweden 8-9 June 2001*.

"Semantic Versioning 2.0.0." `http://semver.org/`.