Phoenix Event Data Set Codebook (0.0.1b)

Caerus Associates, January 12, 2015

Phoenix is a new event data set produced by researchers at Caerus Associates, Penn State University, Parus Analytics, and the University of Texas–Dallas. It provides freely available data on international and subnational political events, generated daily using open-source, customizable software. The community around the dataset is organized as the Open Event Data Alliance.

Software: https://github.com/openeventdata

Open Event Data Alliance: http://openeventdata.org/

Basic Column Descriptions

At its most basic, Phoenix records events consisting of two actors and a directed event. The actor and action information are both encoded according to the CAMEO scheme.[1] This version of Phoenix is comprised of 26 tab-separated columns.

This release of Phoenix should be considered beta. Several additional columns will be added in the future and not not features (including geolocation) are not consistently implemented throughout the period of the dataset.

Version notes and known bugs

In version 0.0.1b, several features are not implemented or partially implemented:

- Geolocation begins in January and is not available before then. The country information in `CountryCode` will be reported as ISO 3 character country codes, but is currently country names. The next release of the data will accurately reflect the variable name and will include geolocation though the entire period.
- A future release will include two additional columns, which will report the human-readable forms of the actors used to generate the codes. These entities will be resolved to consistent forms, based on their entries in the dictionaries. Thus, "ISIS", "the Islamic State", and "ISIL‘" will all be coded as SYRREB and also reported as SourceActorName = "ISIL" for times when the specific group, rather than all Syrian rebel groups, is the focus of inquiry. This addition will obviate the need to modify CAMEO to give specific actors of interest unique codes.
- A future release will include an additional location-related column, which which will include a flag indicating when the reported lat/lons are country centroids and therefore should be excluded from subnational analysis.
- There are several gaps in coverage. Many of these will be filled in as part of a later release.

[1] http://eventdata.parusanalytics.com/cameo.dir/CAMEO.Manual.1.1b3.pdf

Table 1: Column Names and Brief Descriptions

Number	Column Name	Description
1	EventID	Globally unique event ID & data version
2	Date	Date in YYYYMMDD format
3	Year	Year in 4 digit integer
4	Month	Month in 1 or 2 digit integer
5	Day	Day in 1 or 2 digit integer
6	SourceActorFull	All codes for source actor
7	SourceActorEntity	Entity code for source actor (usually country code)
8	SourceActorRole	Primary role code for source actor
9	SourceActorAttribute	Other attributes for source actor
10	TargetActorFull	All codes for target actor
11	TargetActorEntity	Entity code for target actor (usually country code)
12	TargetActorRole	Primary role code for target actor
13	TargetActorAttribute	Other attributes for target actor
14	EventCode	Event code in CAMEO ontology (includes leading zeros)
15	EventRootCode	Event code aggregated to 20 top-level CAMEO (leading zeros)
16	PentaClass	High level score on conflict–cooperation, material–verbal
17	GoldsteinScore	A -10 to 10 conflict-to-cooperation scale
18	Issues	Thematic information based on keywords in the event sentence
19	Lat	Estimated latitude of the event
20	Lon	Estimated longitude of the event
21	LocationName	Finest resolution location of the event
22	StateName	State or province where the event occurred (not supported)
23	CountryCode	Country where the event occurred (ISO 3 char)
24	SentenceID	Story database ID, followed by the zero-indexed sentence number
25	URLs	URL of the scraped story
26	NewsSources	Abbreviated name of news source

Actor Codes

The `Source` actor is the actor doing the action; the `Target` actor is the recipient or object of the action. The Source and Target actors each have four columns:

The `ActorFull` columns include the `ActorEntity`, `ActorRole`, and `ActorAttribute` information strung together, presenting the highest resolution information about the source or target actor.

The `ActorEntity` columns break out the top-level code, which are almost always country codes, but can also be "IMG" for international militarized group, "IGO" for inter(national) governmental organizations, "MNC" for multinational corporations, or "NGO" for nongovernmental organizations. Country codes are in ISO 3166-1 alpha-3 form. Phoenix only includes events that have both source and target entities, which drastically cuts down on noise and mis-codings.

The `ActorRole` describes the primary sub-state identity of the actor, which gives much more context to international events and makes many domestic events intelligible. The role codes are:

Phoenix also includes semi-colon delimited `ActorAttribute` columns, which provides additional information about the actor, including ethnic and religious codes, whether the actor is a societal elite, and whether rebels are identified as insurgents or separatists, among other things. See the CAMEO codebook for more details.

Action Codes

The `EventCode` is the full CAMEO event code, one of around 250 political actions. For instance, the code for "fight with artillery or tanks" is 194. The first two digits indicate which of the 20 top-level CAMEO categories

Table 2: Actor Role Codes

Role Code	Description
GOV	Government
MIL	Military
REB	Rebel
OPP	Opposition
PTY	Political Party
COP	Police and Internal Security
JUD	Judicial Actor
SPY	Intelligence Service
MED	Media
EDU	Educational Actor
BUS	Business
CRM	Criminal
CVL	Civilian

the event falls into (in this case, 19, corresponding to "fight"). Any additional numbers refine the top level code further, in this case to specify the weapons. The `EventCode` column includes leading zeros, so consider importing as a text field or factor rather than numeric/integer. Again, see the CAMEO codebook for details.

`EventRootCode` presents just the event's top-level, 2-digit CAMEO event category, extracted to make filtering and aggregating easier. This column includes leading zeros.

`PentaClass` is an updated version of the QuadClass categories used in previous event data work and based on the work of Duval and Thompson on scaling international relations data.[2] Previous quad-class implementations sliced the CAMEO categories in a linear fashion. This new implementation takes into consideration what the CAMEO categories actually suggest in terms of material or verbal conflict/cooperation. The changes in this version include the creation of a 0 code for CAMEO category 01 ("Make a Statement"), as well as several codes changing quad classes (most significantly, "Protest" is now material conflict rather than verbal conflict). The categories are as follows:

The `GoldsteinScore` reports the event mapped onto a conflict-cooperation scale, with -10 being the most conflictual and 10 being the most cooperative. The score is included because many existing analyses use it. However, the use of a single score to capture both conflict and cooperation interactions has a number of well-documented problems.[3]

Contextual Information

In addition to the actor-action-actor information that is the heart of event data, we are also interested in the context in which the event occurs. This context can be both thematic and geographic.

The `Issues` column contains semicolon-delimited tags that are attached to events based on keyword matching in the article. These words are intended to add context to the events and to allow for filtering of the events.[4] While TABARI had this functionality in the past, it was not implemented in the Levant dataset or in GDELT. Issues information should allow analysts to better understand the context of the events they are examining, as well as allowing more sophisticated searching and subsetting of the data.

[2]Robert D. Duval and William R. Thompson, "Reconsidering the Aggregate Relationship between Size, Economic Development, and Some Types of Foreign Policy Behavior," *American Journal of Political Science*, Vol. 24, No. 3 (Aug., 1980), pp. 511–525

[3]For a discussion on aggregating and scaling event data, see James E. Yonamine, "Working with Event Data: A Guide to Aggregation Choices", http://jayyonamine.com/wp-content/uploads/2012/06/Working-with-Event-Data-A-Guide-to-Aggregation-Choices.pdf

[4]For the exact terms used, see https://github.com/openeventdata/Dictionaries/blob/master/Phoenix.IssueCoding.txt.

Table 3: CAMEO-to-Penta Class

EventRootCode	Description	PentaClass
01	Make Public Statement	0
02	Appeal	0
03	Express Intent to Coop	1
04	Consult	1
05	Engage in Dip Coop	1
06	Engage in Material Coop	2
07	Provide Aid	2
08	Yield	2
09	Investigate	3
10	Demand	3
11	Disapprove	3
12	Reject	3
13	Threaten	3
14	Protest	4
15	Exhibit Force Posture	4
16	Reduce Relations	3
17	Coerce	4
18	Assualt	4
19	Fight	4
20	Use Unconventional Mass Violence	4

The next five columns (`Lat`, `Lon`, `LocationName`, `StateName`, `CountryCode`) contain geolocation information extracted from the sentence the event was coded from. Our geolocation uses MIT's CLIFF geolocation software, which is built on top of Berico Technologies' CLAVIN and geonames.org's gazetteer. Geolocation in this version is only available beginning in January.

Source Information

One of the major improvements of Phoenix over previous event data sets is the degree to which generated events are connected back to the underlying stories reporting the events. Three columns provide this information:

`SentenceID` reports the ID of the story(ies) in our internal Mongo database of stories. The number after the underscore is the sentence in the story (beginning at 0) that the event was extracted from. This column is semicolon-separated.

`URLs` reports all of the original URLs of the news stories containing the event. Previous datasets have reported only one story per event, if they included any at all. Some websites (especially wn.com), change links quickly, meaning that some links will connect to a different story than they did originally. However, most links still work or contain enough information to locate the original story. This column is semicolon-separated.

`NewsSources` reports an abbreviation of the news sources whose story included the event. Phoenix includes events drawn hourly from 450 English language sources, which can be quickly added to at will. The abbreviations are listed in our GitHub page.[5] This provides both transparency on where events are reported and allows analysts and users to include only events from specific sources, if desired. This column is semicolon-separated.

[5] https://github.com/openeventdata/scraper/blob/master/whitelist_urls.csv