Contemporary infrastructure supporting political event data

Philip A. Schrodt, Ph.D.

Parus Analytics LLC and Open Event Data Alliance
Charlottesville, Virginia USA
http://philipschrodt.org
https://github.com/openeventdata/

Presented at the Data Workshop PreView
German Federal Foreign Office, Berlin
16-17 January 2018

PARUS
ANALYTICS

Event Data: Core Innovation

Once calibrated, monitoring and forecasting models based on real-time event data can be run [almost...] entirely without human intervention

- Web-based news feeds provide a rich multi-source flow of political information in real time
- Statistical and machine-learning models can be run and tested automatically, and are 100% transparent

In other words, for the first time in human history we can develop and validate systems which provide real-time measures of political activity without any human intermediaries

Primary point of these comments

Most of the infrastructure required for the automated production of political event data is now available through commercial sources and open-source software developed in other fields: it no longer needs to be developed specifically for event event production.

This dramatically reduces the costs of implementation and experimentation.

WEIS primary categories (ca. 1965)

01	Yield	11	Reject
02	Comment	12	Accuse
03	Consult	13	Protest
04	Approve	14	Deny
05	Promise	15	Demand
06	Grant	16	Warn
07	Reward	17	Threaten
08	Agree	18	Demonstrate
09	Request	19	Reduce Relationship
10	Propose	20	Expel
		21	Seize
		22	Force

Major phases of event data

- 1960s-70s: Original development by Charles McClelland (WEIS; DARPA funding) and Edward Azar (COPDAB; CIA funding?). Focus, then as now, is crisis forecasting.
- 1980s: Various human coding efforts, including Richard Beale's at the U.S. National Security Council, unsuccessfully attempt to get near-real-time coverage from major newspapers
- 1990s: KEDS (Kansas) automated coder; PANDA project (Harvard) extends ontologies to sub-state actions; shift to wire service data
- early 2000s: TABARI and VRA second-generation automated coders; CAMEO ontology developed
- 2007-2011: DARPA ICEWS project
- 2012-present: full-parsing coders from web-based news sources: open source PETRARCH coders and proprietary Raytheon-BBN ACCENT coder

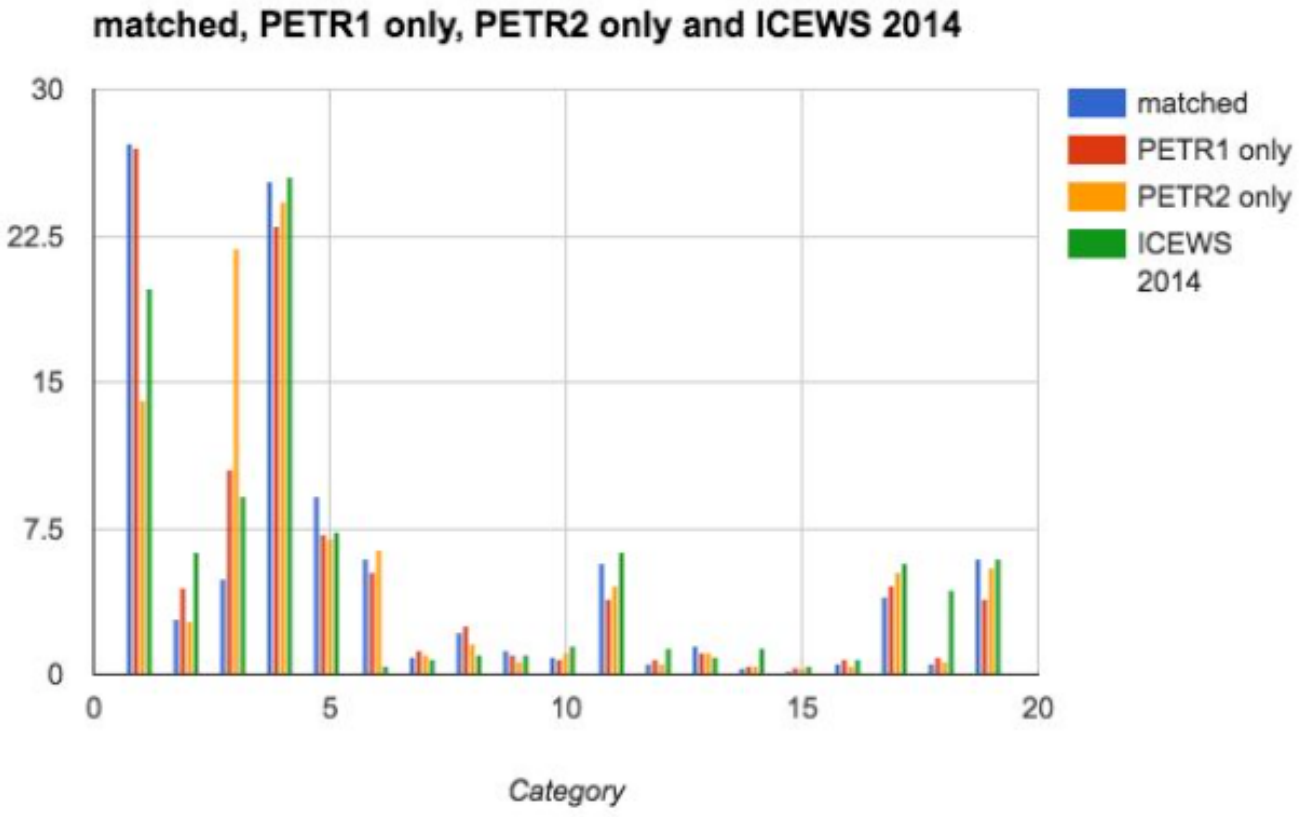
Natural language processing infrastructure

- Named entity recognition is now a standard NLP feature
  - Synonyms can be obtained from JRC
  - Affiliations and temporally-delimited roles can be obtained from Wikipedia
- Parsing, notably through the Stanford CoreNLP suite
  - dependency parsing is very close to an event coding: a basic DP-based coder requires only a couple hundred lines of code `https://github.com/philip-schrodt/mudflat`
- Geolocation `https://github.com/openeventdata/mordecai`
- Robust machine-learning classifiers—SVM, neural networks—as effective filters
- Similarity metrics such as Word2Vec and Sent2Vec for duplicate detection, which also helps error correction
- Machine translation, which may or may not be useful

Event data coding programs

- TABARI: C/C++ using internal shallow parsing.
  `http://eventdata.parusanalytics.com/software.dir/tabari.html`
- JABARI: Java extension of TABARI : alas, abandoned and lost following end of ICEWS research phase
- DARPA ICEWS: Raytheon/BBN ACCENT coder can now be licensed for academic research use
- Open Event Data Alliance: PETRARCH 1/2 coders, Moredcai geolocation. `https://github.com/openeventdata`
- NSF RIDIR Universal-PETRARCH: multi-language coder based on dependency parsing with dictionaries for English, Spanish and Arabic
- Numerous experiments in progress with classifier-based and full-text-based systems

"CAMEO-World" across coders and news sources



Between-category variance is massively greater than the between-coder variance.

Why the convergence?

- This is simply how news is covered (human-coded WEIS data also looked similar)
- The diversity in the language and formatting of stories means no automated coding system can get all of them
- Major differences (PETRARCH-2 on 03; ACCENT on 06, 18) are due to redefinitions or intense dictionary development
- Systems probably have comparable performance on avoiding non-events (95% agreement for PETRARCH 1 and 2)
- Note these are aggregate *proportions*: ACCENT probably has a higher recall rate, but the otherwise pattern is still the same

PLOVER

Political Language Ontology for Verifiable Event Records

Event, Actor and Data Interchange Specification

Open Event Data Alliance

http://openeventdata.org/

http://ploverdata.org/

DRAFT Version: 0.6b2

March 2017

Web infrastructure

- Global real-time news source acquisition and formatting using open-source software
- Relatively inexpensive standardized cloud computing systems rather than dedicated hardware: "cattle" vs "pets"
- Multiple open-source "pipelines" linking all of these components, though these remain somewhat brittle
- ICEWS and Cline Center data sets currently available; Univ. of Oklahoma Lexis-Nexis-based TERRIER (1980-2015) and Univ of Texas/Dallas real-time data should be available soon
- Contemporary "data science" has popularized a number of machine-learning methods that are more appropriate for sequential categorical data than older statistical methods

Remaining challenges: source texts

- Gold standard records
  - These are essential for developing example-based machine-learning systems
  - They would allow the relative strengths of different coding systems to be assessed, which also turns out to be essential for academic computer science publications
  - We don't want "one coder to rule them all": different coders and dictionaries will have different strengths because the source materials are very heterogeneous.

- An open text corpus covering perhaps 2000 to the present. This is useful for
  - Robustness checks of new coding systems
  - Tracking actors who were initially obscure but later become important
  - Tracking new politically-relevant behaviors such as cyber-crime and election hacking

Remaining challenges: institutional

- Absence of a "killer app": we have yet to see a "I've gotta have one of those!" moment.
  - Commercial applications such as Cytora (UK) and Kensho (USA) are still low-key and below-the-radar.
- Sustained funding for professional staff
  - Academic incentive structures are an extremely inefficient and unreliable method for getting well-documented, production-quality software. Sorry.
  - Because they occasionally break for unpredictable reasons, 24/7 real-time systems need to have expert supervision even though they mostly run unattended
  - Updating and quality-control on dictionaries is essential and is best done with long-term (though part-time) staff
  - This effort could easily be geographically decentralized

Thank you

Email:
schrodt735@gmail.com

Slides:
`http://eventdata.parusanalytics.com/presentations.html`

Links to open source software:
`https://github.com/openeventdata/`

ICEWS data:
`https://dataverse.harvard.edu/dataverse/icews`

Cline Center data:
`http://www.clinecenter.illinois.edu/data/event/phoenix/`

Slides from talk summarizing the workshop

[several of these were added after the actual presentation]

What we've seen/learned

- Very large amount of open, near-real-time data is easily available
  - We could, however, probably do more in terms of sharing software
- Extensive analytical tools
- Early warning models are common and may be developing to the point of being a "must have" application
- Monitoring and visualization tools
- Clear international scientific consensus on general characteristics of data and methods
- Easy to incorporate private-sector software development

Open Event Data Alliance software

Birdcage

Basic, Integrated, and Reliably Distributed Coding, Actors, and Geolocation for Events

PETRARCH family of automated event data coders and dictionaries for CAMEO ontology

PLOVER Event Data Ontology

FJOLTYNG: PLOVER- and universal dependency-based event coder

PETRARCH-based web scraping and event coding pipeline

Sources

- International news services: most common sources for most data: quality is fairly uniform but attention varies
- Local media: quality varies widely depending on press independence, local elite control, state censorship and intimidation of reporters
- Local networks: these can provide very high quality information but require extended time and effort to set up
- Social media: notice none of the data projects emphasize these. They can be useful in very short term (probably around 6 to 18 hours) but have a number of issues
  - most content is social rather than political
  - bots of various sorts produce large amount of content
  - difficult to ascertain veracity: someone in Moscow or Ankara may be pretending to be in Aleppo
- not mentioned but available: remote sensing (e.g mapping extent of refugee camps or abandoned farmland)

Is this big data?

Classic definition of "big data": variety, volume, velocity

- Variety: this we have
- Volume: not so much compared to Google, Amazon, medical systems
- Velocity: again, policy-relevant models rarely need true real time, and often use structural data at the nation-year level. Models can be refined and studied, not operated in milleseconds

In addition, we have theories, not just data mining: Amazon [probably] does not have a "theory of backpacks" even if it sells a lot of them. Substantive understanding remains important

The Amazon/Google Theory of Backpacks

Brought to you by Big Data

- If it is August and we have ascertained you are a parent with school-age children, show advertisements for small backpacks
- If it is May and we have ascertained you are between the ages of 18 and 25, show advertisements for large backpacks
- Otherwise show some other advertisement
  - Because I am preparing these slides in Google Docs, I am now seeing ads for SAS's machine-learning software. Seriously. Big Data is Watching You!

Apply this approach to conflict, and I'm guessing Thucydides, Machiavelli and T.R.Gurr still don't have much to worry about

Analytical Challenges

- Data is provided by a large number of small projects with unstable funding: very few institutions delight in funding data collection even while they delight in using data they get for free. Bug? Feature?
- Economic and demographic data, in contrast, is a government function because it is seen as a public good
- Too much data: without a consensus on measures we are wasting a lot of effort on redundant measures
- Too much variety: our data generating processes (and applications) are more heterogeneous than those in most commercial applications
- Importance of transparency and replicability

Do we have *too much* data/variety

WDI has 1500+ indicators available!

Advantages of variety (Kraay)

- Composites have greater stability
- Variance in the measurement provides useful information
- Less affected by biases or methodological weaknesses in individual providers
- Multiple independent sources probably give greater confidence

Do we have *too much* data/variety?

Disadvantages of variety

- Cost and effort
- Some methods—notably the many variants on principal components—for creating composites aren't transparent or unique
- Weak sources introduce noise
- When secondary sources are used to generate the original indicator, those aren't actually independent

Historically, the most robust social science models have used only a small number of easily-measured variables, which is quite a different approach than the current "big data" approach but has a very long record (Kahneman)

Simple models are good!

Recent study on predicting criminal recidivism showed equivalent results could be obtained from

- A proprietary 137-variable black-box system costing $22,000 a year
- Humans recruited from Mechanical Turk and provided with 7 variables
- A two-variable statistical regression model

For this problem, there is a widely-recognized "speed limit" on predictive accuracy of around 70% and, as with conflict forecasting, multiple methods can achieve this.

Source: *Science* 359:6373 19 Jan 2018, pg. 263; the original research is reported in *Science Advances* 10.1126/sciadv.aao5580 (2018)

PITF operational modeling approach

- Accumulate a large number of variables from open sources and exhaustively explore combinations of these using a variety of statistical and machine-learning approaches: this establishes the out-of-sample "speed limit"
- The "speed limit" should be similar to the accuracy of human "super-forecasters" (Tetlock)
- Construct operational models with "speed limit" performance using very simple sets of variables—typically fewer than five—using the most robustly measured of the relevant independent variables

Simple models are transparent; robust measures are transparent and inexpensive

Challenges applying this to foreign policy

- Integrating quantitative analysis into traditionally qualitative decision-making
- Economic historians have found that efficiently integrating a new technology (e.g. steam power; electricity; computers) into an industry takes about 20 years, a human generation
- Rare events and probability analysis are difficult for everyone, including statisticians (Kahneman)
  - Questions such as the relationship between climate change and conflict are *very* difficult to study and we won't have immediate answers
- Visualization is also difficult (Tufte): machine-assisted self-deception
- Political sensitivity: transparency might help here

Supplementary Slides

PLOVER objectives

- Only the 2-digit event "cue categories" have been retained from CAMEO. These are defined in greater detail than they were in WEIS and CAMEO.

- Some additional consolidation of CAMEO codes, and a new category for criminal behavior

- Standard optional fields have been defined for some categories, and the "target" is optional in some categories.

- A set of standardized names ("fields") for line-delimited JSON (`http://www.json.org/`) records are specified for both the core event data fields and for extended information such as geolocation and extracted texts;

- We have converted all of the examples in the CAMEO manual to an initial set of English-language "gold standard records" for validation purposes—these files are at `https://github.com/openeventdata/PLOVER/blob/master/PLOVER_GSR_CAMEO.txt`—and we expect to both expand this corpus and extend it to at least Spanish and Arabic cases.

Event, Mode, and Context

Most of the detail found in the 3- and 4-digit categories of CAMEO is now found in the *mode* and *context* fields in PLOVER. More generally, PLOVER takes the general purpose "events" of CAMEO (as well as the earlier WEIS, IDEA and COPDAB ontologies) and splits these into "$event - mode - context$" which generally corresponds to "$what - how - why$." We anticipate at least four advantages to this:

1. The "$what - how - why$" components are now distinct, whereas various CAMEO subcategories inconsistently used the *how* and *why* to distinguish between subcategories.
2. We are probably increasing the ability of automated classifiers—as distinct from parser/coders—to assign *mode* and *context* compared to their ability to assign subcategories.
3. In initial experiments, it appears this approach is *much* easier for humans to code than the hierarchical structure of CAMEO because a human coder can hold most of the relevant categories in working memory (well, that and a few tables easily displayed on a screen)
4. Because the words used in differentiate *mode* and *context* are generally very basic, translations of the coding protocols into languages other than English is likely to be easier than translating the subcategory descriptions found in CAMEO.

PLOVER output


```
{
    "id": "test-0056-0036_1",
    "date": "2015-02-12",
    "source": [{"actorText": "Russian Foreign Minister Sergei Lavrov", "code": "RUS", "sector": "GOV"},
               {"actorText": "Iranian counterpart Mohammad Javad Zarif", "code": "IRN"}],
    "target": [{"actorText": "Syria crisis", "code": "SYR"}],
    "event": "DISCUSS",
    "eventText": "discussed",
    "mode": "mode-holder",
    "context": "context-holder",
    "text": "MOSCOW: Russian Foreign Minister Sergei Lavrov and his Iranian counterpart Mohammad Javad
     Zarif discussed the Syria crisis by phone Wednesday, the Russian Foreign Ministry said in a statement
    "language": "en",
    "publication": "mudflat test data",
    "coder": "Parus Analytics",
    "version": "0.5b1",
    "dateCoded": "2017-03-20",
    "comment": "test output from mudflat",
},
```

PLOVER: ASSAULT modes

Name	Content
beat	physically assault
torture	torture
execute	judicially-sanctioned execution
sexual	sexual violence
assassinate	targeted assassinations with any weapon
primitive	primitive weapons: fire, edged weapons, rocks, farm implements
firearms	rifles, pistols, light machine guns
explosives	any explosive not incorporated in a heavy weapon: mines, IEDS, car b
suicide-attack	individual and vehicular suicide attacks
heavy-weapons	crew-served weapons
other	other modes

Adapted from Political Instability Task Force Atrocities Database:
`http://eventdata.parusanalytics.com/data.dir/atrocities.html`

PLOVER: general contexts

Name	Content
political	political contexts not covered by any of the more specific categories below
military	military, including military assistance
economic	trade, finance and economic development
diplomatic	diplomacy
resource	territory and natural resources
culture	cultural and educational exchange
disease	disease outbreaks and epidemics
disaster	natural disaster
refugee	refugees and forced migration
legal	national and international law, including human rights
terrorism	terrorism
government	governmental issues other than elections and legislative
election	elections and campaigns
legislative	legislative debate, parliamentary coalition formation
cbrn	chemical, biological, radiation, and nuclear attacks
cyber	cyber attacks and crime
historical	event is historical
hypothetical	event is hypothetical