# Keynote: Current Open Questions for Operational Event Data

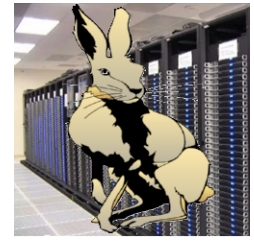
Philip A. Schrodt, Ph.D.

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

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

AESPEN @ LREC 2020
9-11 June 2020

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

# But there are some. . . well. . . "issues". . .


Source: Clionadh Raleigh. Keynote: Too soon? The limitations of AI for event data. AESPEN @ LREC 2020, 9 June 2020

# 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

# 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 |

# PLOVER categories (ca. 2018)

**Agree**
**Consult**
**Support**
**Cooperate**
**Aid**
**Concede**
**Retreat**

**Investigate**
**Crime**

**Demand**
**Disapprove**
**Reject**
**Threaten**
**Protest**
**Mobilize**
**Sanction**
**Coerce**

**Assault**
**Fight**

# Some non-conflict, non-protest domains that could be addressed with suitably broad event data

- Illiberal democratic transitions:
  Turkey, Hungary, Poland, Hong Kong, maybe USA
- 21st century right-wing populist movements—Tea Party (USA), Brexit (UK), Yellow vest (France)—vs. 1920s-30s right-wing populism in Italy, Germany, Spain, UK, USA
- Diversionary hypotheses
  - External: Countries take advantage of international crises to escalate in other domains (1956: simultaneous occurrence of Suez crisis and Soviet invasion of Hungary)
  - Internal: Countries escalate in international relations to distract from internal problems (Falklands/Malvinas 1982; Russian annexation of Crimea 2014)

Event data would not be used exclusively here but would be a useful and inexpensive (if already coded) supplement to other measures such as opinion polling and economic data.

# 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"
- Contemporary "data science" has popularized a number of machine-learning methods that are more appropriate for sequential categorical data than older statistical methods
- Multiple open-source "pipelines" linking all of these components, though these remain somewhat brittle

# 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`
- Word embeddings such as BERT and ELMO deal with the issue of synonyms, a major weakness in older dictionaries
- Similarity metrics such as doc2Vec 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`

As reported in this workshop, numerous experiments are currently in progress using machine learning rather than parsers, rules, and patterns

# Open Event Data Alliance software

**Birdcage**

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

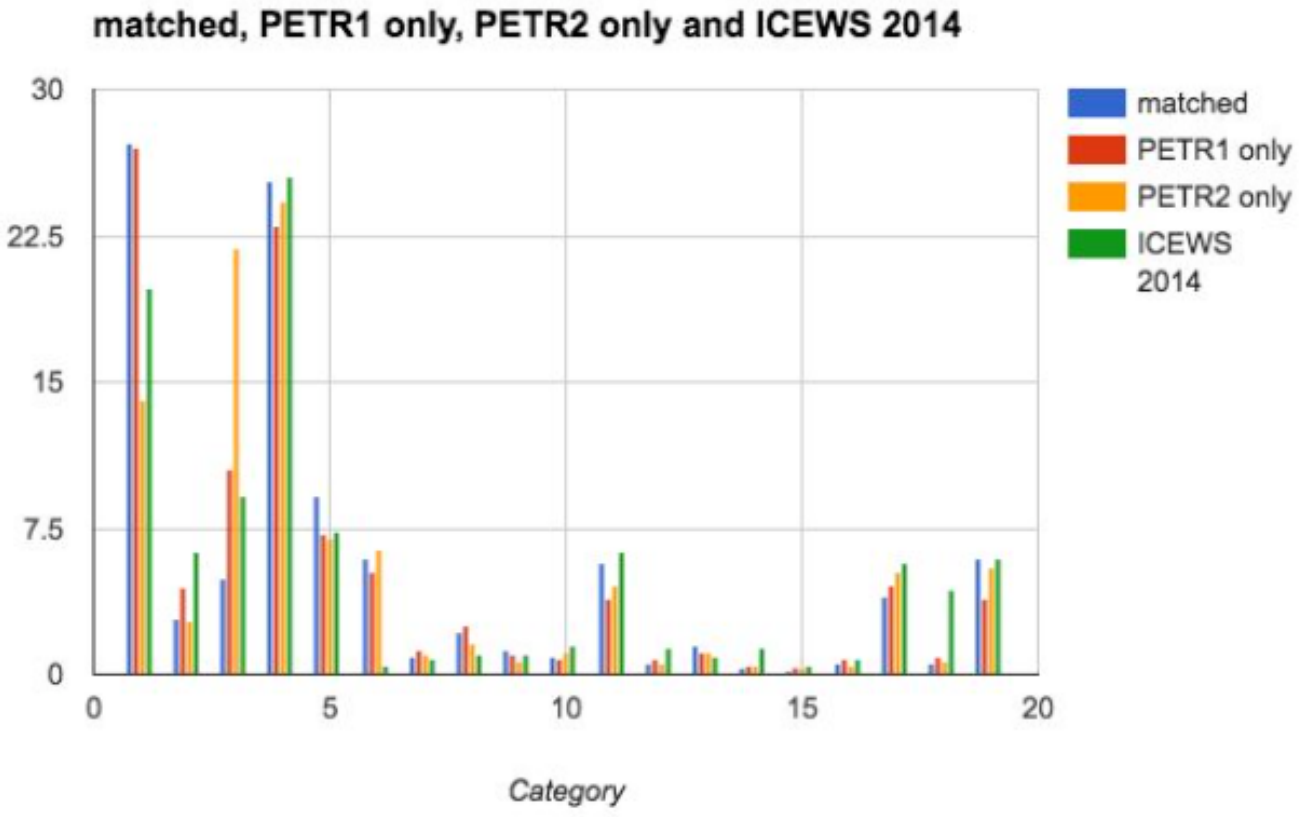
PETRARCH family of automated event data coders and dictionaries for CAMEO ontology

PLOVER Event Data Ontology


FJOLTYNG: PLOVER- and universal dependency-based event coder


PETRARCH-based web scraping and event coding pipeline

# "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 overall pattern is still the same

# Seven contemporary challenges (roughly in order of priority)

1. Near-real-time systems with fully transparent pipelines
2. Systematic experimentation and comparison of parser-rule-pattern coders and example-based coders
3. Open benchmarks and, ideally, training cases
4. Replace the CAMEO system with a more general system encompassing, e.g. disaster and disease, legislative, electoral, and criminal behavior
5. Develop and test multiple duplicate detection systems, ideally based in texts, not the coded events
6. Ascertain the value-added from hundreds of localized sources: this may vary by regions
7. Systematic experimentation and comparison of native-language coders vs coders using machine translation

# 1. Near-real-time systems with open and transparent pipelines

"Open and transparent" means *every* automated component of the data generating process—acquisition and filtering pipeline, coders, and both actor and event dictionaries (or training cases)—needs to be available with no restrictions on experimentation.

As Clionadh pointed out, updates on the order of a couple days to a week are sufficient for operational purposes, hence "near-real-time." Actual "real time" reports are extraordinarily noisy and quickly require revision.

# Known weaknesses in existing systems: ICEWS

- Coder and event dictionaries are proprietary
- Development of coder ended around 2015; non-commercial license for coder, while free, has restrictions on experimentation
- Open actor dictionaries do not cover most coded actors, so additional proprietary dictionaries must be in use somewhere in the pipeline (e.g. nation names)
- Very ad hoc geolocation based on a few heuristics
- Coverage is still biased to Asia, the original focus of the 2008-2011 DARPA ICEWS program
- South Asia, particularly India, is massively over-represented (though if you want to study India, ICEWS is very nice: go for it!)
- Almost no coverage of China (source issues) and USA (restrictions due to funding)

# Known weaknesses in existing systems: PETRARCH/Phoenix family

- Entire PETRARCH family of open-source coders is experimental and was never refined to an operational level
- PETRARCH-2—coder used in Phoenix databases—was a [brillant] summer internship project that was never fully completed
  - Code contains assorted unimplemented features, e.g. internal `pico` code modification system is just a proof-of-concept with rules for a single verb, ABANDON
  - Verb dictionaries are *radically* modified from their TABARI base but have never been curated
  - Actor dictionaries have not been systematically updated

Pro-tip for computer scientists: Just because code is available on GitHub doesn't mean it should be deployed in operational settings.

# Operational systems?

This is likely to require 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 pipelines 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 and shared between NGOs and academics

# 2. Parser-rule-pattern coders vs example-based machine-learning coders

# Parser-rule-pattern coders

Advantages

- A well-established and "good enough" technology: first refereed publications were in 1994
- Multiple open source parsers, coders, and [CAMEO] dictionaries
- Open dictionaries do not have intellectual property issues

Disadvantages

- Dictionary development is incredibly labor intensive
- Probably too brittle to work with machine-translated texts; very few languages (just Spanish, really) are covered beyond English
- This is a 30-year-old and highly specialized approach: except for parsers and NER, no work outside the very small, if thoroughly international, event coding community

# Example-based machine learning

Advantages

- Major innovations in NLP in recent years using a variety of open and easily implemented neural-network-based approaches
- Human and rule-based coding is nowhere close to 100% accurate, so there is room for fairly sloppy approaches
- "Tin standard" (vs "gold standard") examples are much cheaper to generate than linguistic rules and patterns
- Probably will be more effective on machine-translated texts

Disadvantages

- Unproven
- A large number of training cases may be required
- Training cases will have intellectual property issues unless abstract or synthetic cases can be used

# 3. Open benchmarks and training cases

- Intellectual property issues are obviously central here. Commercially-generated information does not want to be free.
- If an ontology like CAMEO—roughly 300 categories in a complex hierarchy—is fully calibrated, a very large number of benchmark cases is required. Systems with less complexity will, of course, require fewer resources.
- Computer scientists will not work on anything—or at least can't publish anything—without established benchmarks
- The issue is primarily the event dictionaries and categorization (along with target identification): actor dictionaries are a relatively simple NER problem which we can supplement with resources such as JRC Names and Wikipedia. NER disambiguation—Tripoli, Lebanon vs. Tripoli, Libya (ICEWS hasn't figured this one out. . . )—remains an open problem but affects a relatively small number of cases.

# Three possible approaches

1. Develop something similar to—or gain access to—the DARPA-funded Linguistics Data Consortium "Gigaword" news article corpus. However, Gigaword itself has two problems
   - It is not easy to license or access outside a few academic institutions
   - The ca. 2000 corpus is increasingly dated

2. Develop synthetic cases, perhaps using GANs, which can be used to train systems which work on real news reports. TABARI's "Lord of the Rings" validation suite (250 cases) were manually-generated synthetic cases free of IP issues.

3. Develop abstracted cases which qualify as facts—which cannot be copyrighted under US law—rather than creative content, though there might be significant differences between US and European law on this

# Example of an abstracted parse

```
"text": "Malaysia's king met parliamentarians for a second day on
        Wednesday to try to end political turmoil by finding someone
        able to form a government or by calling a new election after
        Mahathir Mohamad's shock resignation as prime minister."

"fjmlParse": [
      "ACTOR-0", "meet-VERB", "AGENT-0", "for-ADP", "a-DET", "second-ADJ",
      "day-NOUN", "on-ADP", "PROPN-0", "try-VERB", "end-VERB",
      "political-ADJ", "turmoil-NOUN", "by-ADP", "find-VERB", "someone-NOUN",
      "able-ADJ", "form-VERB", "a-DET", "AGENT-1", "or-CCONJ", "by-ADP",
      "call-VERB", "a-DET", "new-ADJ", "election-NOUN", "after-ADP",
      "PROPN-1", "shock-NOUN", "resignation-NOUN", "as-ADP", "AGENT-1"
    ],
```

Abstraction (via `spaCy`)

- All proper noun phrases (NER) and agents (dictionary-based) are replaced with placeholders
- Verbs and nouns are replaced with lemmas (e.g. “met”→“meet”, “to try”→“try”)
- parts-of-speech tags added to tokens
- Punctuation removed

# 4. CAMEO was never intended as a general-purpose coding ontology and has significant limitations

# 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: 0.7b1

March 2020

# 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.
- (event, mode, context) triples replace the hierarchical coding structure of WEIS and CAMEO; event-dependent "mode" and "context" can probably be coded using simple classifiers (SVM or neural networks)
- Some additional consolidation of CAMEO codes, and a new category for criminal behavior
- All of the examples in the CAMEO manual have been converted to an initial set of English-language "gold standard records, " though these are probably not sufficient for full calibration and certainly not for training.
- A set of field names for JSON records are specified for both the core event data fields and for extended information such as geolocation, citation, and parsing; getting a standard data-interchange format would facilitate the development of open-source utilities and JSON is much more human-friendly than `.csv` or `.xlsx`.

# 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",
},
```

# ICEWS output

```
Event ID        Event Date      Source Name     Source Sectors  Source Country  Event Text
Intensity       Target Name     Target Sectors  Target Country  Story ID        Sentence Number
Publisher       Source  Headline        City    District        Province        Country Latitude
Longitude       productID       holdingID
35288929        2017-01-13      Government (Malawi)     Government      Malawi  Occupy territory
-9.5    Malawi  NULL    Malawi  51071520        1       Other Publisher (OSC)   OSC     Malawi: MRA
seals Times Group offices   Blantyre        Blantyre District       Southern Region Malawi  -15.785
35.0085 6c15e214        ffac-4bff-a678-cfc30e106bcd
35290881        2020-03-09      Lawyer/Attorney (Mali)  Social,Legal    Mali    Criticize or denounce
-2      Ministry (Mali) Government      Mali    51076719
5       Other Publisher (OSC)   OSC     Malian Magistrates Denounce 'Intolerable Interference' by
Government in Judiciary       NULL    NULL    NULL    Mali    12.65   -8      35565297
8ec3-4e0d-a9c6-ea1b9576e473
35285530        2017-04-12      Aleksandr Shokhin       Education,Business
NGOs,Social,Nongovernmental Organizations / Activists       Russian Federation      Engage in
symbolic act  0       Business (Russia)       Business,Social Russian Federation      51062438
6       Other Publisher (OSC)   OSC     Russia: How Atomic Closed Cities Retooled, Survived Market
Conditions (2017)    NULL    NULL    NULL    Russian Federation      55.7522 37.6156 4bd14a7a
011b-48ca-a02e-d92eea252971
```

# 5. Duplicate detection

- In ICEWS, major events can have 100 to 200 duplicates. This will create large and systematic biases in any models based on the data.
- While duplication conveys some information in terms of importance, it is mostly a function of whether the event occurred somewhere with large numbers of journalists (and, sometimes but not always, nice hotels)
- Certain types of events—"developing stories," "when it bleeds it leads"—consistently generate more duplicates. "Celebrity events" such as Tiananmen Square, the 9/11/2001 attacks, and the Chibok kidnappings generate thousands of repetitions over years.
- One-a-day de-duplication based on coded events amplifies coding errors
- News story clustering is an established technology: EMM, Google, several papers presented here

# Possible alternative approach

1. Code a *set* of events from sentences extracted from a cluster of related stories

2. Retain only those events which are coded from several stories

3. Record the number of stories and number of sources for each event

In other words, instead of coded-event-based deduplication—"one-a-day" filtering—deduplication based on clusters of related stories, which is much closer to how human analysts and coders process duplicate reports.

# 6. How many sources are really needed?

During the period 2005 to 2015, essentially all news sources in the world established some on-line presence.

- International news services: these are the most common sources for most data; the quality is fairly uniform but attention varies and generally focuses on areas and issues of interest to people who have a lot of money
- Local media: quality varies widely depending on reporting styles, 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. There are also ethical issues/risks when governments or IGOs do this: it is an NGO option.

# More sources may not be better for general event data

- Predictably, most coded events are generated by a small number of sources: of the roughly 1000 sources used to generate ICEWS, the top 20 sources generate 82% of unique events; top 50 generate 92%
- Variations in local styles vs. the "pyramid' international style. More generally, do the local sources differ significantly from the training sets, and have they been systematically tested?: establishing levels of coding accuracy in 1000 sources is very expensive
- (Dis)Information controlled by local elites and/or manipulated by state-level entities
- No evidence to date that event data can find highly localized "needle in a haystack" events having predictive value

Note: ICEWS calculations exclude India, which is vastly oversampled in ICEWS, and included articles from the US-government Open Source Center

# 7. Native-language coders vs machine translation

- Investments in commercial machine translation will be orders of magnitude greater than anything we can apply to event coding
- Advantage: Machine translation has been improving rapidly and past weaknesses of pattern-based coders may be less of a problem with newer translation technology
- Disadvantage: Texts describing political events are often quite different than texts typically used to train translation systems, and politics tends to use a lot of language-specific idioms
- If machine-earning systems become common, a hybrid approach of curating translated texts and then using the original texts to train a language-specific system may work

# Final thoughts

- We should not have “one data set to rule them all”
- Follow the approach of hurricane and snowstorm forecasters who triangulate results of multiple independently developed data sets and models which have different assumptions and strengths

# "What would this look like if it were easy?"—Tim Ferris

Table: Cost to put one person in Earth orbit, 2019 $US-millions

| Project | Development | Per-person cost |
|---|---|---|
| NASA Space Shuttle | $27,400 | $170 |
| NASA Orion (projected) | $23,600 | $291 |
| SpaceX CrewDragon | $ 1,700 | $ 55 |

*Economist*, 4 June 2020

Possibly apochryphal:
Around 2014, Google assessed the performance of deep neural networks combined with word embeddings in machine translation projects, and decided to discard 500,000 lines of hand-crafted code developed over a decade, replacing this with 500 lines of Tensorflow calls.

# 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`

# Supplementary Slides

# PLOVER: 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: 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 |