# Seven observations on the newly released ICEWS data

Philip A. Schrodt, Ph.D.
Parus Analytics LLC
Charlottesville, VA 22901

schrodt735@gmail.com

Full URL

https://asecondmouse.wordpress.com/2015/03/30/seven-observations-on-the-newly-released-icews-data/

Posted on March 30, 2015

# Seven observations on the newly released ICEWS data

Posted on March 30, 2015

Before we get to the topic of the post, the usual set of apologies about the absence of recent postings—starting with that Duke Nukem Forever style "Feral+well, whatever." Isn't that I've dropped out, it is rather that I've been too busy with other projects. And by the way, I haven't retired,[1] unless logging 2,200 hours of work last calendar year is "retired." Someday, things will slow down, I'll get to the backlog. But enough about me…that's not what we're up to today.

Instead, the point of today's posting is to comment upon the long, long awaited release of a public version of the Integrated Conflict Early Warning System (ICEWS) dataset, which appeared without fanfare on Dataverse late in the afternoon on Friday, 27 March, with Jay Ulfelder probably responsible for first spotting it.

This is a massive resource: The investment in the ICEWS project, albeit not all of the funding going into the data, is probably roughly equalled the whole of NSF spending on all international relations and comparative politics research during the time it was active. As with any large data set it is going to take a while to figure out all of its quirks. The ever-resourceful David Masad already has some excellent instructions and initial analyses and visualizations here, and I'm sure more will be forthcoming in coming weeks—and years—but I wanted to use the occasion to alert my [ever declining] readership to this, and provide some initial observations.

## 1. It exists!

Long overdue, to be sure, given that at the ICEWS "Kick-off Meeting" in 2007 we were assured that everything in the project would be open, a concept almost immediately quashed, and then some, by the prime contractors. I'm pretty sure we have the persistent and unrelenting efforts of Mike Ward and Philippe Loustaunau to thank for the release. I've also got a pretty good idea of who is responsible for the delays but, well, let's just focus on positive things right now.[2]

Here's the link to the data: http://thedata.harvard.edu/dvn/dv/icews. There are actually four "studies" involved

- 28075: This is the main data set: 26 files, most of these are around 30 Mb each. Took me a couple tries to get some of them, though that may have been due to a lousy wireless connection on my end. It's Dataverse; it will work.
- 28117: Aggregated data: this may the quickest way to get into the data, assuming what you

are interested in is covered by one of the very large number of aggregations that have already been computed. I've not really looked at these yet, and much of the documentation appears oriented to a proprietary dashboard which is not provided, but particularly for people not comfortable working with very large disaggregated data sets, it could be very useful.

- 28118:These are the dictionaries, more on this below.
- 27119: This was the big disappointment: we had been told the release would include a set of "gold standard cases", which we assumed would be the much-needed gold standard cases needed to validate event coding systems, but these, alas, are just some sort of esoteric records associated with the much-disputed ICEWS "events of interest." Hard to imagine it will be much use for anyone, but I've been wrong before.

## 2. Dictionaries!

As we've [3] been arguing for decades, the primary advantage of automated coding is the ability to maintain consistent coding across data sets being coded across a number of years and, through the dictionaries, to have a high level of transparency.[4] For that you need the dictionaries as well as the coder, and the KEDS project [5] has been consistently providing those as part of the data. To its great credit, ICEWS has followed this norm, and what dictionaries they are!: a primary actor dictionary with over 100,000 political actors. The format is derivative of that used in TABARI and PETRARCH—I'm guessing it will take about fifteen minutes to write a converter.

The agent dictionary—also provided, along with a somewhat cryptic "sectors" dictionary— on the other hand, is definitely a work in progress, though probably fine for the major sub-national actors, which for the most part are still those established by Doug and Joe Bond's work on PANDA and IDEA back in the 1990s, and subsequently incorporated intoCAMEO. The quirk that really sticks out for me is the treatment of religion: for ICEWS, it seems "Christian" and "Catholic" are separate primary categories [6]—granted, the late Ian Paisley [7] would agree—and the Great Schism of 1054 apparently was no big deal. The whole of Judaism gets only two entries, rather an oversimplification for the neck of the woods I'm usually studying. There is an extraordinarily eclectic set of ethnic groups—with a distinct oversampling in India—and, well, overall, the agent dictionary is sort of like rummaging through some old trunk in your grandparent's attic [8], and I'm pretty sure we're well ahead of this at the Open Event Data Alliance.

ICEWS did not provide event code dictionaries, which are presumably tightly linked with the proprietary BBN ACCENT coder. This is a bit of an issue, since ACCENT does not actually code CAMEO, but their own variant which they have documented in a very extensive manual. Not

ideal but no worse than the situation with any human-coded data.

## 3. You'll need to convert it for statistical analysis but I've got a program for that.

The public-release ICEWS uses a very quirky format that is apparently designed to be read rather than analyzed, as the underlying codes are presented in verbose, English-language equivalents. Unless you are ready to settle into a few quiet evenings reading through the 5-million records, you'll probably want to use the data in statistical analyses, which means you'll want to get shorter codes. It just so happens, I've got an open-source program for that at https://github.com/philip-schrodt/text_to_CAMEO and I've even provided you'all with COW codes as well as ISO-3166-alpha3 codes. I didn't fully convert the sector dictionaries, but this will at least give you a good start.

## 4. Massive use of local sources

That old criticism that event data are nothing but the world as viewed from the point of Western imperialists? This will be hard to sustain with ICEWS, which uses hundreds of local sources, and each event contains information on the source. I've only looked in detail at 2013, and here these follow more or less a rank-size distribution, with some of the major international sources (Xinhua, BBC) being major contributors, but the tail of that distribution is extremely long.

## 5. The distribution is flat

While the internet, and new social media more generally, are revolutionizing our ability to inexpensively generate large-scale datasets relevant to the study of political behavior, a serious problem has been dealing with the exogenous effects of the rapid expansion of internet-based sources that began in the mid-2000s. Any "dumpster diving the web" approach leads to an exponential increase starting about this time, which for any statistical analysis is a bug, not a feature.

ICEWS avoids this: they seem to be using a relatively fixed set of sources, and the total density is largely flat. As Masad's visualizations and some others I've seen show, there appears to be a bit of variation—1995 and 1996 seem undersampled—and more will probably appear as further research is done, since there have been major changes in the international journalism environment beyond just the increase in the availability of reports, but these variations are not

exponential, and can probably be accommodated with relatively simple statistical adjustments.

## 6. 80% precision, but no assessment of the accuracy

The release is accompanied by an extensive analysis showing that the "accuracy" of the ACCENT coder is around 80%. Which would be very nice, except that the study actually assesses not accuracy, but *precision*, which, while interesting, gives us no information whatsoever on the measure most people are interested in: the probability of correctly coding a randomly chosen sentence (accuracy), rather than the probability that a sentence that was coded was coded correctly (precision). Echoing the exchange between Col. Harry Summers and one of his Vietnamese counterparts over the unbroken string of US battlefield successes, the assessed precision "May be so, but it is also irrelevant."

The arguments here are a bit technical, though involve nothing more than simple algebra, so I've relegated this to an appendix to this post. The upshot, to paraphrase Ray Stevens, "Yo selected on the dependent variable, and I can hear yo' mama sayin', "You in a heap o' trouble son, now just look what you've done""

## 7. It should splice with Phoenix

The current dataset has a one-year embargo, though the word on the street is that the embargo will remain at just one year, more or less . That is, the data will be periodically updated, ideally monthly, perhaps quarterly. [Addendum: in a very promising sign, the March 2014 data were indeed made available on 1 April 2015.] This will be adequate for most retrospective studies, but still won't help with the real-time forecasting that event data are increasingly used for.

Here the recently-released OEDA Phoenix data set comes to the rescue, or will once we've got another four or five months of ICEWS data, as Phoenix gets going around the beginning of July 2014. Provided ICEWS is updated regularly, within a fairly short period of time one should be able to use the ICEWS 1995-2014 data for calibration, and then use Phoenix to cover the end of ICEWS to the present (Phoenix is updated daily).

Assuming, of course, that the data are sufficiently similar that they can be spliced, possibly with some adjustments. The major distinction between the data sets is likely to be the sources, with ICEWS using Open Source Center feeds and proprietary data services, and Phoenix a white-list of Web-based sources. This is likely to make a big difference in some areas—in the very limited exploration I've done, ICEWS seems to disproportionately focus on India, for example, and for

statutory reasons, contains no internal data on the US—and less in others. Actor dictionaries will not be an issue as the ICEWS dictionaries could be used to code the Phoenix sources, though this may not be necessary.

The different coding engines may or may not make a difference: in the absence of a confirmable set of gold standard cases for events, and verb dictionaries, we will need a significant period when the two sets overlap to find out whether the two systems perform significantly differently. My guess is that they won't differ all that much, particularly if common actor dictionaries are used, since that both coders are based on full parsing, and the differing sources will be the bigger issue. Both Phoenix and ICEWS provide information on the publications where the coded text came from, so these could be filtered to get similar source sets.

In the absence of a public version of ICEWS overlapping with the [still relatively brief] Phoenix data, we can only do indirect measures of the likely similarities, but some quick analyses I've done comparing marginals of the first six months of Phoenix with the last six months of ICEWS indicate two promising points of convergence: the density of data (events per day) was quite similar and—even more telling—the marginal densities of the event types were very similar (actors less so but again, that's easily corrected since the ICEWS actor dictionaries are public). Again, we won't be able to do the more crucial test—the correlation of dyad-level event counts—until there is a substantial overlap in the public data, but initial indications are promising.

# What needs to be done (all open)

Call me a greedy anti-intellectual knuckle-dragging Neanderthal—and you will—but when I read a recent article in *Science* about the construction of an esoteric scientific instrumentwhose construction cost was $300-million and annual operating costs are $30-million, and then compared that with the pittance that is being allocated—when we can avoid our programs being shut down altogether [9]—for event data which could contribute significantly to at the very least to increasing the ability of NGOs to accurately anticipate situations where "bad things might happen" [10], or even to a reality-based foreign policy, I get a tad irritated. Consider these aspects of the instrument in question:

- it may not work—its also-costly predecessor did not—and half of the project is situated in a place in Louisiana that makes it less likely to work, suggesting it is largely a mindless boondoogle. A boondoogle located in Louisiana, I'm shocked, shocked.
- if it does work, it merely further confirms a century-old theory which we've got complete confidence in already, and as the *Science* article points out, is confirmed billions of times

each day, for example as a smart phone displays inappropriate content having determined that you are a male walking within fifty meters of a Victoria's Secret outlet store. [14]

- and the predictions of the theory at issue were already confirmed by other observational evidence four decades ago, for which the discoverers got a nice trip to Stockholm.

Which is to say, this is just the natural sciences equivalent of a performance art project [13], but at a rather higher price tag. And unlike space telescopes and Mars rovers, we don't even get nice pictures from it.

So, like, if we can spend what will probably eventually total some half-billion dollars before this thing winds down, presumably with the yawn-inducing equivalent of the umteenth iteration of "Hey, ya'know, Mars once had water on it!!" how about spending 1%—just a lousy 1%—of the that amount (which is probably also about 10% of the cost of ICEWS) on enhancing event data? And this time with social scientists in charge, not folks whose prime competence is raiding the public purse under the guise of protecting our national interests against opponents who disappeared decades ago. Oh, and every single line and file of the project open source. I'm just asking for 1%! A guy can dream, right?

So, say we've got $5-million. Here's my list

1. **Open gold standard cases.** Do it right: the baseline will be the openly available Linguistic Data Consortium GigaWord news files, use a realistically large set of coders with documented training protocols and inter-coder performance evaluation, do accuracy assessments, not just precision assessments. Sustained human coder performance is typically about 6 events per hour —probably faster on true negatives—and we will need at least 10,000 gold standard cases, double-coded, which comes to a nice even $50K for coders at $15/hour, double this amount for management, training and indirects, and we're still at only $100K.

2. Solve—or at least improve upon—the **open source geocoding** issue. This is going to be the most expensive piece, and could easily absorb half the funds available. But the payoffs would be huge and apply in a wide number of domains, not just event data. I'd put $2M into this.

3. **Extend CAMEO and standard sub-state actor codes**, using open collaboration among assorted stakeholders with input from various coding groups working in related domains. We know, for example, that one of the main things missing in CAMEO are routine democratic processes such as elections, parliamentary coalition formation, and legislative debate, and there are people who know how to do this better than us bombs-and-bullets types. On sub-state actor

coding, religious and ethnic groups are particularly important. I'm guessing one could usefully spend $250K here. Also call it something other than CAMEO.

4. **Automated verb phrase recognition and extraction**, which will be needed for extending the CAMEO successor ontology. I actually think we're pretty close to solving this already, and we could get some really good software for $50K. [11] If that software works as well as I hope it will, then spend another $250K getting verb-phrase dictionaries for the new comprehensive system.

5. **Event-specific coding modules**, for example for coding protests and electoral demonstrations. Open-ended, but one could get a couple templates for $100K.

6. Systematic **assessment of the native language versus machine translation**issue. That is, do we need coding systems (coders and dictionaries) specific to languages other than English, particularly French, Spanish, Arabic and Chinese [12], or is machine-translation now sufficient—remember, we're just coding events, not analyzing poetry or political manifestos—so given finite resources, we would be better off continuing the software development in English (perhaps with source-language-specific enhancement for the quirks of machine translation). Hard to price this one but it is really important so I'd allocate $500K to it

7. Insert **your favorite additional requirements** here: we've still got $1.75M remaining in our budget, which also allows a fair amount of slack for excessively optimistic estimates on the other parts of the project. Or if no one has better ideas, next on my list would be systematically exploring splicing and other multiple-data-set methods such as multiple systems estimation. And persuade Lockheed to dust off the unjustly maligned JABARI—or make the code open source if they have no further use for it—and give us another alternative sequence based on that program.

All this for only 1% of the cost of a single natural science performance art project! Come on, someone out there with access to the public trough—or even some New Gilded Age gadzillionaire—let's go for it! Pretty please?

# Footnotes

1. Yeah, I can just imagine the conversations at ISA in New Orleans (I was on Maui. Just on vacation. Really.)

> "Hey, Schrodt really disappeared once he left Penn State. Figured that would happen…"
>
> "Really, it's bad: I heard that he was last seen on the side of the exit ramp off I-99 to Tyrone, looking really gaunt and holding a cardboard sign that said 'Will analyze mass atrocities for food'."
>
> "Yes, that's right: so sad. So keep that in mind if you are thinking about leaving academia, or even imaging the possibility of asking any senior faculty to get their fat Boomer butts out of the way."

Well, no, that's not really accurate. But we'll save that for another blog entry. Meanwhile, you can follow me on GitHub. And I'll be at EPSA in Vienna.

2. And keep our faith in the wheel of karma.

3. I'm not exactly sure who "we" is—I'm neither royalty nor, to my knowledge, have a tapeworm—but I'm trying to represent the views of a loose amalgam of people who have been working with machine-coded event data for a good quarter-century now.

4. Total transparency when the coding software is available, which is not the case here, but even without the software these dictionaries are a huge improvement over the transparency in most human coding projects, where too many decisions rest on an undocumented and ever-shifting lore known only to the coders.

5. Or whatever it should be called: it will always be KEDS—Kansas Event Data System—to me.

6. Two Protestant denominations get designations at the same level as "Herdswoman" and "Pirate Party"—Episcopal (but not Anglican) and Methodist—and there is an entry for "Maronite." That's it: no Lutherans, no Baptists, no Pentecostals, no Mormons, not even the ever-afflicted Jehovah's Witnesses. In fact in the ICEWS agent ontology, the only religions worthy of subcategories are Christian, Catholic, Buddhist, Hindu and Moslem, though the latter has not been affected by that unpleasantness at Karbala in 680 CE. The ontology developers, however, appear to have spent a bit too much time watching re-runs of the Kung-Fu television series—or more ambiguously, Batman Begins—as only Buddhism produces "warriors."

7. To say nothing of the late Fred Phelps.

8. Yeah, yeah, they moved to a condo in Arizona two decades ago and the old place was torn

down and replaced with a MacMansion, but it still makes for a nice metaphor.

9. Though I did notice that Senator Jeff Flake was one of the few Republicans not to throw his lot in with the GOP efforts to provide free policy advice to the Islamic Republic of Iran, so perhaps his M.A. in Political Science did some good.

10.I think we are now at a point where these things can make a serious difference: The absence of major electoral violence in the 2013 Kenyan elections and—fingers crossed—the 2015 Burundi elections may eventually be seen as breakthroughs on this issue.

11. But meanwhile, don't get me started on the vast amounts that is wasted onhiring programmers who never finish the job. Really, people, the $75 to $150 an hour to hire someone with a professional track record who will actually write the programs you need is a better deal than spending $25,000+ a semester—stipend, tuition and indirects, and this is actually a low estimate for many private institutions—for one or more GRAs are supposed to be learning programming but who, in fact, stand a pretty good chance of getting absolutely nowhere because writing sophisticated research software does not, in many instance, provide a good pedagogical platform. No matter what your Office for the Suppression of Research says.

12. Hindu/Urdu is also important in terms of the number of speakers, but, for better or worse, the media elites in the region use English extensively.

13. If you aren't familiar with this concept, Google "bad performance art." NSFW.

14. To clarify, not the precise example used by *Science*. Continue reading →

Posted in Methodology | 4 Comments | Edit