Twenty Years of the Kansas Event Data System Project

Philip A. Schrodt
Dept. of Political Science
University of Kansas
Blake Hall
Lawrence, KS 66045
785.864.9024 (phone) - 785.864.5700 (fax)
Email: schrodt@ku.edu

VERSION 1.0: June 12, 2006
An extended version of this article is available at
http://www.ku.edu/∼keds/KEDS.history.html

1 In the beginning there was event data

Event data—sequences of nominal codes recording the interactions among political actors—have been a major focus of quantitative international relations (IR) research since the 1960s and 1970s. Until recently most event data analysis used either Edward Azar's (1982) Conflict and Peace Data Bank (COPDAB) or Charles McClelland's (1976) World Event Interaction Survey (WEIS), but over the past decade the combination of machine-readable news reports and automated coding have dramatically reduced the costs of generating new data sets.

This article will discuss the development of the automated coding systems of the Kansas Event Data System (KEDS) project, which developed the first automated coding system—the eponymous KEDS computer program—that could produce data acceptable in refereed articles. This is a personal history rather than an attempt to provide a fully balanced account, and is told from my perspective with an emphasis on the developments that seemed most important to me; other recollections may differ.

2 Roots

The initial enthusiasm for event data among researchers was tempered by the fact that it was very expensive to produce. Historically, event data have been coded by legions of bored students flipping through copies of *The New York Times.* Event data collection slowed in the late 1970s as funding—which had largely come from the U.S. Department of Defense Advanced Research Projects Agency (DARPA)—was discontinued. The existing data sets continued to be widely used—one study found them to be the third most frequently used form of data in quantitative IR research—but they were not updated.

In the mid-1980s, I became involved in a fascinating set of work influenced by the then very trendy research on "artificial intelligence" (AI). A number of political scientists, dissatisfied with the limitations of conventional statistical analysis, attempted to apply AI to the formal study of international relations and foreign policy. This produced a variety of quite novel approaches using computational methods before being tragically destroyed after exposure to a particularly virulent strain of post-modern deconstructionism, for which at the time there was very little acquired immunity.[1] My own efforts at computational modeling largely revolved around applications of event data, in particular trying to use these to simulate the problem of reasoning by analogy: see `http://www.ku.edu/∼keds/papers.dir/Schrodt.PRL.2.0.pdf`.

I used an assortment of event data sets in this work, mostly versions of COPDAB and WEIS that I had acquired from various sources. It was clear, however, that these were not very dense, particularly on the Arab-Israeli conflict, the area where I was also doing field research. More critically, they were not being updated, so I could not use them to study contemporary issues or do true forecasting.

In the process of doing some contract programming for a West Coast consulting firm developing a political decision-support system "for a major U.S. ally", I became acquainted with (and obtained a copy of) a WEIS data set that had been collected by various defense consulting firms for no less a client than the National Security Council in the Reagan White House, where event data was championed by a McClelland student named Richard Beale. This effort continued until Beale's untimely death in 1985.

One of the novel features of this version of WEIS was the inclusion of brief English-language summaries of the news story that generated the event. This provided an opportunity to check whether it was possible, in principle, to go from a natural language text to event codes. Working with a Northwestern University undergraduate, David Leibsohn, I developed a simple computer program that mapped keywords to event codes, and presented this at the 1985 International Studies Association meetings (Schrodt and Leibsohn 1985). This produced credible results, and in particular provided early evidence that while some WEIS categories were subtle and difficult to code consistently (probably for humans as well as machines), many of the most common events—notably meetings and uses of force—were straightforward because they were described using a very distinct and specific vocabulary.

In the late 1980's, the National Science Foundation undertook a major initiative titled "Data Development in International Relations" to update the most widely used international relations data sets (and, one suspects, reduce the grousing from the quantitative IR community over the resources going to the American National Election Survey). The second phase of DDIR was headed by my dissertation adviser Dina Zinnes and her colleague, the late Richard Merritt (Merritt, Muncaster and Zinnes 1994). A group of about twenty researchers was convened, and eventually NSF invested about $350,000 in a number of different event data projects.

Automated coding was seen as a possible, but by no means proven, approach to reducing the costs of producing event data. My initial contribution to DDIR was a machine-learning program-that was an elaboration of the Schrodt-Leibsohn work. It worked reasonably, though not spectacularly, well, and its ultimate contribution was to simply provide some of the basic code for what would develop into Keds. The machine-learning aspect was consistent with most of my work in AI, but turned out to be a dead-end: Keds and Tabari both eventually required extensive, and highly expert, dictionary development by humans. In retrospect, this simply reflected a general lesson from automated natural language processing in the 1980s—humans are so good at language, and language is such an idiosyncratic human construct, that it is better to let humans tell a machine what to do (and then have the machine routinely do it) than to try to develop machine learning algorithms.[2]

By the late 1980s, we had a reasonable set of techniques that provides credible results, but we had only shown that we could map natural language news *summaries* into event categories. This was a long way from the "holy grail" of producing data directly from news wire accounts. The closest thing available seemed to be various machine-readable indices, but these typically had insufficient information to resolve events beyond the level of the dozen or so major "cue categories"

[1] Sylvan and Chan (1984) provides early examples of the AI approach; Hudson (1991) has later, more sophisticated examples, and Trappl (2006) shows the approach is not dead yet.

[2] The contemporaneous DARPA sponsored "Message Understanding Conference" project (DARPA 1993) had somewhat similar objectives—extracting details of terrorist incidents from news wire stories (and involving real computer scientists and real linguists!)—and came to similar conclusions: systems using extensive phrase dictionaries developed by humans far out-performed machine-learning systems.

in the event coding schemes.

At this point I had a chance encounter with a student involved with the University of Kansas (KU) debate program who, on hearing about my research, asked "Why don't you just use Reuters?—it's available on NEXIS." "Where do I get NEXIS?" "At the Law School—all the debaters use it there."[3]

I arranged a meeting with one of the Law School librarians, who demonstrated that one could, in fact, download Reuters stories via a dial-up connection. He assured us that there was no marginal cost to the Law School for using the service, and then noted that the Law library was closed overnight. He then suggested we wait while he showed us a couple more things about the system, but he had to check with NEXIS technical support first. He placed a couple of calls, asking questions at the level of "Which one is the *any* key??" but in the process of getting authorization, repeated, loudly and slowly "Our NEXIS password is. . . ".

Dial-up connection, library is closed, here's the password: we can take a hint. We set up a simple script to automate a log-in to NEXIS from about 2 a.m. to 5 a.m., and over the next several months downloaded tens of thousands of stories.

3 KEDS

Based on the WINR demonstration, DDIR provided a small $40,000 grant in 1991-1993 for the development of what became the KEDS program. While I continued to do most of the computer programming, the bulk of the value-added from the KEDS project has been provided by the twenty or so "dictionary developers" who have been involved with the project and devoted thousands of hours to refining the dictionaries that are essential to producing data.[4] As I am endowed with the inter-personal skills typical of a computer programmer, this aspect of the project has been directed by my collaborator Deborah J. Gerner.

While the KEDS work for DDIR included some experimentation with German-language sources and foreign policy chronologies (Gerner et al 1994), most of our development focused on WEIS coding of interactions in the Middle East reported by the Reuters news service in English. We focused on this area both because it is very thoroughly covered in the international press, but also because we were doing field work in the area and could therefore cross-check the validity of event data based on our experience in the region.

KEDS had its professional debut at the 1992 International Studies Association meetings in Atlanta. The conference paper was being written, typically, at almost the last minute, and focused on a 12-year time series for the Arab-Israeli conflict. A number of different pieces had to come together to generate this—downloading and formatting the Reuters stories, on-going dictionary development, and aggregation of the resulting events into an interval-level time series using the Goldstein (1992) scale—so only when the paper was nearly finished that could I actually look at the results. I still vividly remember finally getting the Israel-Palestinian series, and plugging it into MS-EXCEL to get a basic plot. My great fear was that the Palestinian intifada would not show up in the data. To my tremendous relief, there it was as a lovely (if noisy) spike followed by an exponential decay, the most conspicuous feature of the series. On early-1990s hardware, the system coded about 70 events per second, a huge improvement over human coding projects, which typically have a sustained output of five to ten events per coder per hour.

One of the people who heard that ISA presentation was Doug Bond from the Program on Nonviolent Sanctions in Conflict and Defense at the Center for International Affairs at Harvard who was beginning the development of a new event coding scheme, the Protocol for the Assessment of Nonviolent Direct Action (PANDA). The PANDA project worked in close collaboration with us

[3] Yes, kiddies, in those days NEXIS was a highly restricted resource, not something available on a browser at most research universities. We also walked to and from school every day in the snow. Barefoot. In June. Uphill. Both directions.

[4] An eloquent description of the challenges of dictionary development can be found in Joseph Pull's "Ode on Coding" `http://www.ku.edu/∼keds/home.dir.ode.html` which Pull wrote prior to leaving the project for Yale Law School.

for the next two years during the most intense development of Keds in dictionary development, identification of bugs, and validation.

The PANDA work eventually spun off a commercial event-coding operation—VRA, Inc. [`http://vra.com`]—which developed a coding program that used quite different principles than Keds . The PANDA coding system, with the added collaboration of Craig Jenkins (Ohio State) and Charles Taylor (VPI) morphed into the Integrated Data for Events Analysis (IDEA) coding scheme (Bond et al 1997). Reuters reports dealing with the entire world have been coded for by VRA for 1985-2004; the resulting data set contains about 10-million events and can be downloaded from `http://gking.harvard.edu/data.shtml`

4 Tabari

Keds was written in the Pascal programming language and worked only on Apple Macintosh computers. The choice of Pascal made sense at the time—it was the core language for the Macintosh operating system and my visceral loathing of Microsoft made the Macintosh the only option if I were to be doing the programming.[5] However, by the late 1990s Pascal had been largely superceded by the C/C++ as the most common general-purpose programming language and compiler support for the language was dwindling. Furthermore, while Keds was generally stable from 1995 to 2000, it contained some deep-seated idiosyncrasies that could only be eliminated by completely re-writing the program.

In response to this, Tabari —Textual Analysis By Augmented Replacement Instructions—was created in the spring of 2000. It is based on the same sparse-parsing principles as Keds but is written as "open-source" code in ANSI C++ and was immediately ported to the Linux and Windows operating systems. The conversion to C++ resulted in a program that was substantially faster than the Pascal code—the program codes about 8,500 records per second even on an inexpensive machine, a $500 1.2 Ghz G4 Mac Mini. This is about 300-times faster than Keds, and about 33-million times faster than typical human coding. A simple keyboard-driven interface is implemented using the Unix "ncurses" terminal library, and consequently we now have 100% compatibility between the Macintosh and Unix/Linux versions, as well as allowing the program to run remotely from a server.[6]

The most recent development in our project has been the CAMEO—Conflict and Management Event Observations—coding scheme. This is a new coding system specifically designed for automated coding, and has also evolved to accommodate the post-Cold War emphasis on political events involving sub-state actors. This work has been primarily done by Deborah Gerner and her graduate student Ömür Yilmaz.

In addition to Tabari and CAMEO, the KEDS project has produced an increasingly diverse set of utility programs to support the production of event data. The most important of these have been our automated downloading programs, which have evolved from a script running a dial-in connection followed by processing in Pascal to an integrated perl program that does downloading and reformatting from HTML files taken off the web. Actor_Filter is another one of our stalwarts: this identifies the actors in a set of text based on capitalization patterns, and produces a keyword-in-context index of these, sorted by frequency.

5 Funding

The Keds and Tabari systems—both the programming and the more labor intensive dictionary development—have been funded by a combination of NSF grants and government contracts, with occasional bridge funding from KU, and an interesting contract doing conflict monitoring for a

[5] Linux was still a gleam in Linus Torvald's eye when work began on Keds, and machines running various flavors of Unix were quite expensive. This is Kansas: we don't do expensive.

[6] We've had less success finding someone to keep the Windows version current. Whatever. . .

Swiss-based NGO. We've been lucky (okay, we've also made our luck...): money has always been available for the tasks we needed to do, and in fact at times we've turned down work because of our limited number of trained coders [take the hint: if you can master the relevant tools, there is more funding available for this than we can handle].

We've always kept our work unclassified, for both principled and pragmatic reasons. We've no desire to become the Kansas equivalent of Los Alamos scientist Wen Ho Lee and advance the career of a zealous FBI agent anxious to get out of Topeka.[7] Meanwhile the value-added of classified material in the real world doesn't quite live up to its portrayal in the movies: consider for example the timely warnings provided by classified analysis on the collapse of the Soviet Union, India nuclear weapons tests, Iraqi WMDs and the recent Hamas electoral victory. As the "open source" concept was popularized in the late 1990's, we shifted all of our work to that mode.

The KEDS project has generally been a relatively small affair: typically Gerner and me, one or two graduate assistants, a data manager, and a half-dozen coders. We've been larger—last summer PRI's accountant came to me and said "Do you realize you've got twenty people on your payroll??" (uh, no, I hadn't—kinda creeps up on you...)—but smaller is the norm. I occasionally get emails from people wanting to come and visit "our shop"—presumably envisioning the vast KEDS Building with its own cafeteria, weight room and day-care center. I explain that there really isn't a shop, just a web page: nothing to see here, move along, move along.

6 Mama don't your babies grow up to be event data analysts

At this point we have spent five years in initial experimentation with automated coding methods, devoted about fifteen years to operational program and dictionary development, produced regional data sets for about thirty countries, and now have the capability of maintaining data sets with a resolution of about a day at close to zero marginal cost (or at least a lower marginal cost than any other known method of creating data in the social sciences, including curb-stoning). Event data analysis has therefore taken the quantitative international relations world by storm, right?

Well, no. While articles utilizing event data have appeared on a relatively regular basis in all of the refereed "sacred journals" that carry quantitative work, it remains very much a niche approach in international relations and comparative politics. Individuals focusing on event data analysis have, with a couple of exceptions, not fared particularly well in the academic job market: in fact the individual who I feel was doing some of the very best work outside of KU was, at last report, running a coffee shop.

Event data has fared substantially better in the policy community, and several people who have been unable to secure academic employment have gone on to positions as quantitative policy analysts in the defense and intelligence communities. There they pull down salaries twice those of academics, don't have to attend faculty meetings, don't grade bluebooks, and can't take their work home because it is classified.

This latter point, however, means that we have had very little feedback from the policy community. Six months after one of my best-trained students took a defense-related job where he was hired explicitly for his event data training, we met at the APSA. "Job going well?" says I. "Yep." says he. "Bet you can't tell me a single thing about it." says I. "Yep." says he.[8]

Two things stand in the way of this. The first is paradigmatic: quantitative research in international relations is dominated by the "Correlates of War" approach that has almost nothing in common with event data analysis. COW studies typically involve the analysis of interval-level variables measured at the nation-state dyad-year level across two centuries and the entire international system. In contrast, contemporary event data analysis focuses nominal measurements in protracted conflicts across a couple of decades or less, but with daily resolution and an increasing focus on

[7] Trust us, any FBI agent based in Topeka will want to get out.

[8] This "I could tell you but then I'd have to kill you" problem has also affected forecasting models using rational choice methods. These may, or may not, be extensively used in the intelligence community, depending on who you want to believe.

sub-state actors. The COW community has generally focused on retrospective inference guided by theoretical issues; the event data community on policy-relevant forecasting.

The second problem involves the shortage of nominal-level time series methods. These exist—for example hidden Markov models—but they are generally closer to pattern recognition methods than to classical frequentist statistics. Interval-level time series are used extensively in econometrics, a field already familiar to most political methodologist. In contrast, the two major sources of nominal-level methods are the very unfamiliar fields of linguistics and bioinfomatics.

Central to the Japanese "manufacturing miracle" of the second half of the twentieth century was the concept of *kaizen*—incremental improvement. A worker's suggestion that increases the quality of a product by only 0.1% will, when combined with similar suggestions by thousands of workers over a period of decades, provide the technological leverage to reduce a device for playing music from the size of a suitcase to the size of a pocket knife, while hugely increasing capacity and quality.

KEDS was an improvement over human coding. TABARI and the VRA CODER are improvements over KEDS, and TABARI can be incrementally augmented through the open-source development process. The CAMEO and IDEA coding schemes are improvements over WEIS and COPDAB. Each time a coder finds another verb phrase to add to the dictionary, or adds another name to the list of actors being coded, the probability of sentences being coded correctly increases, however slightly.

At this point we probably have a good idea of how to produce event data—all event data articles published in major journals over the past ten years have used machine-coded data, and the last major human coding project was shut down in 2004 following a comparison between its data and a comparable data set produced using TABARI. We still need to take the next step in figuring out some really good things to do with it. But at least we've started.

7 For further information:

The KEDS project maintains a very extensive web site at `http://www.ku.edu/∼keds` At this site you will find the most recent versions of the software and documentation, assorted coding dictionaries, data sets and utility programs, a FAQ (frequently-asked-questions) section, and copies of papers from our project and related efforts.

8 references

Azar, Edward E. 1980. "The Conflict and Peace Data Bank (COPDAB) Project." *Journal of Conflict Resolution* 24:143-152.

Bond, Doug, J. Craig Jenkins, Charles L. Taylor and Kurt Schock. 1997. Mapping Mass Political Conflict and Civil Society: The Automated Development of Event Data. *Journal of Conflict Resolution* 41, 4: 553-579.

Defense Advanced Research Projects Agency. 1993. *Proceedings of the Fifth Message Understanding Conference (MUC-5).* Los Altos,CA: Morgan Kaufmann.

Gerner, Deborah J., Philip A. Schrodt, Ronald A. Francisco, and Judith L. Weddle. 1994. "The Machine Coding of Events from Regional and International Sources," *International Studies Quarterly* 38:91-119.

Goldstein, Joshua S. 1992. "A Conflict-Cooperation Scale for WEIS Events Data." *Journal of Conflict Resolution* 36: 369-385.

Hudson, Valerie, ed. 1991. *Artificial Intelligence and International Politics.* Boulder: Westview

McClelland, Charles A. 1976. *World Event/Interaction Survey Codebook.* (ICPSR 5211). Ann Arbor: Inter-University Consortium for Political and Social Research.

Merritt, Richard L., Robert G. Muncaster and Dina A. Zinnes, eds. 1993. *International Event Data Developments: DDIR Phase II.* Ann Arbor: University of Michigan Press.

Schrodt, Philip A. and David Leibsohn. 1985. "An Algorithm for the Classification of WEIS Events from WEIS Textual Data." Paper presented at the International Studies Association, Washington, March 1985.

Schrodt, Philip A. and Deborah J. Gerner. 1994. "Validity assessment of a machine-coded event data set for the Middle East, 1982-1992." *American Journal of Political Science* 38: 825-854.

Trappl, Robert, ed. 2006. *Programming for Peace : Computer-Aided Methods for International Conflict Resolution and Prevention.* Berlin: Springer.