

Integrated Data for Events Analysis (IDEA): An Event Typology for Automated Events Data Development*

DOUG BOND, JOE BOND, CHURL OH

Program on Nonviolent Sanctions and Cultural Survival, Harvard University

J. CRAIG JENKINS

Mershon Center for International Security, Ohio State University

CHARLES LEWIS TAYLOR

Department of Political Science, Virginia Polytechnic Institute and State University

This article outlines the basic parameters and current status of the Integrated Data for Event Analysis (IDEA) project. IDEA provides a comprehensive events framework for the analysis of international interactions by supplementing the event forms from all earlier projects with new event forms needed to monitor contemporary trends in civil and interstate politics. It uses a more flexible multi-levelled event and actor/target hierarchy that can be expanded to incorporate new event forms and actors/targets, and adds dimensions that can be employed to construct indicators for early warning and assessing conflict escalation. IDEA is currently being used in the automated coding of news reports (Reuters Business Briefs) and, in collaboration with other projects, in the analysis of field reports. The article summarizes the conceptual framework being used in this data development effort, its major variables, and its geographic and temporal coverage.

Introduction

Event analysis has a long, rich history in international conflict research but, in the past few decades, it has been bypassed in favor of simpler methods focusing on general conditions (e.g. the presence of armed conflict) and institutional standards (e.g.

human rights protections). This has been due to two problems: (1) the difficulty of generating large amounts of high-quality data; and (2) limitations in traditional events frameworks, which have had an inflexible structure and lacked analytic dimensions that could be used for early warning and assessing conflict escalation. The first problem has been addressed by the development of automated coding through such systems as the Kansas Events Data System (KEDS), its successor TABARI (Textual Analysis By Augmented Replacement

* A revised version of a paper originally presented at Uppsala University, Sweden, 8–9 June 2001. See <http://www.pcr.uu.se>. The authors gratefully acknowledge the collegial support the KEDS/TABARI group generously offered throughout our long and fruitful collaboration. Correspondence: dbond@wcfa.harvard.edu.

Instructions), and the VRA[®] Knowledge Manager. What in the past took months or years to code can now be done in a matter of weeks with coding reliability that is comparable to human coders (Gerner et al., 1994; Schrodtt & Gerner, 1994; King & Lowe, 2003; Jenkins, Abbott & Taylor, 2002). This article addresses the second problem – the limitation of traditional event frameworks. We outline a synthetic framework for international event analysis – IDEA (Integrated Data for Event Analysis) – outline its conceptual structure and major variables, and discuss current data development that is using this framework. The IDEA framework is available on the VRA website (<http://vranet.com/IDEA>) and can be expanded to incorporate additional event forms and actors (sources and targets). It also contains summary indicators, such as the coerciveness and contentiousness of events and conflict-carrying capacity (Jenkins & Bond, 2001) that can be used to gauge conflict escalation. We begin by discussing the problems with existing event frameworks and how IDEA builds on PANDA (Protocol on Nonviolent Direct Action [Bond & Bond, 1995]), WEIS (World Events Interaction Survey [McClelland, 1978]), and the political events data of the *World Handbook of Political and Social Indicators* (or *World Handbook* [Russett et al., 1964; Taylor & Hudson, 1972; Taylor & Jodice, 1983]).

International Event Frameworks: Problems and Prospects

The major problem with existing event frameworks is their lack of summary measures for capturing conflict escalation. Traditionally conceived as an unranked series of discrete event forms for describing relations, WEIS has the virtue of flexibility and greater breadth than alternative frameworks but lacked summary dimensions for gauging conflict escalation. It also lacked

actor and target coding, which was a virtue insofar as this advanced the idea of event forms independent of specific actors, but was a limitation in analysis. To create conflict dimensions, analysts have typically scaled WEIS events using Goldstein's (1992) conflict/cooperation weights. When the PANDA project began adapting the WEIS scheme to capture intrastate events, it became apparent that new event forms (e.g. protest demonstrations) would have to be added. It was also evident that it would be useful to gauge the dimensions of coerciveness and contentiousness as well as physical violence to construct summary indicators of conflict processes, such as conflict-carrying capacity.

In its original formulation, the concept of conflict-carrying capacity (Bond & Vogeley, 1995; Bond et al., 1997) was expressed as the proportion of direct action multiplied by the proportion of forceful action subtracted from one. This approach provided the desired interaction effect between contentiousness and violence, but at the cost of conceptual simplicity and empirical imprecision. In our second iteration (Jenkins & Bond, 2001) of the conflict-carrying capacity measure, we separated civil challenges from governmental repression to better pinpoint the source of instability. While WEIS and other event frameworks provided the raw material for the contentiousness, coerciveness, and violence dimensions in terms of events, the dimensions were not inherent in the framework per se.

The major virtue of the WEIS scheme was its two-level hierarchy of 'cue' and more specific events, which made it more flexible than a single list of discrete events. Another virtue was focusing on events that could be related to news and other reports of the 'who did what to whom, where, and how' framework of event research. Other international events frameworks, such as COPDAB (Conflict and Peace Data Bank [Azar, 1980]) and MID (Militarized Disputes [Jones,

Bremer & Singer, 1996]), mix events with general statements of condition (e.g. full-scale war). A third virtue is rejecting the assumption that events are consistently ordered from 'conflict' to 'cooperation', which should instead be scaled by analysts for particular purposes (McClelland, 1983). The IDEA framework has maintained these principles while expanding the event framework as outlined below. It is useful to briefly summarize the history of the projects leading directly to IDEA.

PANDA

The PANDA project (Bond & Bond, 1995) began in 1988 as an attempt to systematically assess the incidence and impact of non-violent struggle throughout the world. It has continued now for over 14 years at the Weatherhead Center for International Affairs, sponsored by the Program on Non-violent Sanctions through 1994 and thereafter by its successor, the Program on Nonviolent Sanctions and Cultural Survival. The original purpose was to determine under what conditions contemporary nonviolent struggle anywhere in the world had been successful in effecting social, political, or economic change, or in resisting tyranny. To the extent that nonviolent struggle was found, evidence was also sought to determine whether this form of 'people power' was spreading.

After a pilot study based on human 'hand coding' of global news reports, the project searched for automated tools to facilitate its research. For five years, the PANDA team worked with the KEDS (now TABARI) software (see <http://www.ukans.edu/~keds/index.html>). Several lessons became clear as we began to assess global news reports of nonviolent struggle. First, nonviolent direct action, no less than violent direct action, was reported in abundance, even by mainstream news media. Second, nonviolent direct action, like its violent counterpart, was

variable in its outcomes, with the strategic performance of protagonists playing a pivotal role. Third, the tradition of human coding of voluminous electronic news reports posed technical as well as conceptual research challenges, particularly with respect to the unit and level of analysis.

The World Handbook

The three editions of the *World Handbook* pioneered the coding of domestic political event data for most countries of the world. Indicators included measures of both peaceful and violent events of mass political protest, sanctions by governments, armed civil conflict, and changes of government executives. It has been almost two decades since the publication of the last *World Handbook*, and this type of cross-national event research has virtually disappeared from the literature. In its place, conflict analysts have either focused more narrowly on events in specific countries and time periods or used more simple 'conditions' measures, such as the presence of armed conflict (e.g. Eriksson, Wallensteen & Sollenberg, 2003; Esty et al., 1998) and violations of human rights standards (e.g. Henderson, 1991; Poe & Tate, 1994). Policymakers have lacked a timely empirical basis for comprehensively assessing civil and international conflict.

The automated coding of global news reports makes it possible once again to create large and comprehensive international event datasets. We are currently constructing a successor to the events data component of the *World Handbooks* from the intrastate events coded with the IDEA protocol.

The IDEA Framework

IDEA is designed to include all the event forms, actors, and targets of these earlier events frameworks. By using a four-level event hierarchy, IDEA can include new event forms as specifications of more general event

forms. At the higher levels, events are defined independent of specific actors and targets, making the framework more flexible. In its current form, IDEA includes nearly all the event forms from WEIS, PANDA, *World Handbook*, CAMEO (Gerner et al., 2002), and MID.¹ IDEA is also explicitly designed to support the automated coding of text. The event hierarchy means that coding errors typically fall into the same general event category and can more easily be corrected, and that new refinements in event forms (e.g. 'suicide bombings', which constitute a newly evolved type of 'armed action') can be added at the terminal or fourth event level. Terminal event forms are those that have no subforms.

Automated Data Development

Owing to the large costs and logistic problems of human coding, most of the above-mentioned events datasets are not continuously updated, and event analysts have focused on limited time periods and territories. The long time-lag between events and their availability to policy analysts (often several years) has undermined the use of events data research as a policy tool. The development of automated coding makes feasible the development of large-scale event datasets on a near real-time basis, suitable for policy as well as academic analysis.

The IDEA protocol and the VRA[®] Knowledge Manager software system operate together to automatically generate social, economic, environmental, and political events data and to display them in summary form in terms of event counts and various scales. Past work has often focused on the simple counts of particular types of events but, following work on international interactions (Goldstein, 1992; Schrodtt & Gerner, 2000; Goldstein & Pevehouse,

1996), we think summary indices are often more telling and reliable. While each record in the event data matrix constitutes an individual event report, the overall contour of a conflict or struggle is too often lost in the details. Indeed, we view the coded events as *input* for an analyst whose major concern is assessing the overall trend. By summarizing these event matrices in tables, graphs, and maps constructed from event counts, the analyst can quickly gauge the trend of events in an ongoing situation. As peaks and troughs become apparent, the VRA[®] Knowledge Manager is programmed to allow the analyst to 'drill' down to review the underlying reports that generated the anomalous data-point in question. Thus, the system is designed to illuminate trends in near real-time and to help analysts gain an understanding of conflict at a glance, while also providing for close-grained analyses of specific event sequences and turning points.

Given this capability for automated monitoring of an ongoing situation from both global news feeds and field situation reports,² custom datasets can now be generated at will. To presage an argument made below, this 'data on demand' approach better facilitates the incorporation of ongoing improvements in measurement and offers data more appropriate to specific research questions. These custom datasets are dynamic in that they can be modified on demand with any number of variations in the coding rules or term definitions, and

¹ For the cross-mappings of IDEA to/from WEIS, *World Handbook*, MID, and CAMEO, see <http://vranet.com/idea/>.

² We are working with several IO and NGO groups on a web-based data-entry tool to manage security incidents and to do field situation (baseline) reporting. Since the input formats for field and news media reports are the same, we can triangulate the 'view from above' (an international news agency) with the 'view from below' (field-based IO/NGO staff). An example of a customized field reporting system using the IDEA framework is the FAST project conducted by the Swiss Peace Foundation (<http://www.swisspeace.ch>). This project uses trained field reporters to recount events occurring in Central and South Asia, the Balkans, and the Horn of Africa.

across a wide range of substantive applications. These datasets are tailored to the user's concerns and can incorporate revisions as needed. Since automated coding using the IDEA protocol is transparent and consistently applied, analysts can revise it and conduct further tests on the same input to determine the effects of adjustments. This data-on-demand approach shifts our attention from the fixed 'one size fits all' datasets of the past to the tools used to develop custom sets as needed.

VRA[®] Knowledge Manager has three components: the parsing; the field reporting; and the display modules. The automated parser receives input text in the form of some defined interface and breaks it up into parts of speech like nouns, verbs, and attributes and, in a procedure akin to diagramming sentences, discerns meaning from semantic and syntactical structure. The parser draws upon both syntactical rules and semantic relations to assign meanings to classes of words, making it superior to pattern recognition methods relying on discrete literal words. It handles large volumes of text and orders it into the appropriate syntactical and semantic units, and then associates them with appropriate event codes. The parser's output matrix of 'events' – *who does what to whom, when, where, and how* – can then be analyzed by visual, statistical, and other means. Below, we provide an outline of the variables currently used in the system, but first we provide a brief discussion of the unit of analysis. In the following discussion, we draw on our experience coding Reuters Business Briefs but, in principle, the VRA[®] Reader can be applied to any English-language text with consistent style and grammar.

Unit of Analysis

Syntactically, the unit of analysis for the Reader is the independent clause; that is, the Reader identifies discrete event reports

comprised of a subject and predicate, even if the agent of the subject is implied. For example, 'a bomb went off in London today' carries an implied but unidentified agent that placed the bomb. For most purposes, the source and target are required, so the system's effective base unit of analysis may be usefully characterized as a report of *who does what with/to whom*, or as Schrodt & Gerner (2001) put it, an event is a clause 'with a transitive verb'.

In the bomb explosion example, the clause-bound unit of analysis is congruent with what humans do when coding events data. However, most contentious politics events are more commonly considered at a higher level of aggregation by human coders. For example, humans typically think of 'protest demonstration' as taking place on a certain day in a certain location. Analysts typically bound events by a 24-hour clock and require that the event have a city-day location. Human coding thus often diverges from the machine's strict clause-bound unit. Human coders also often consult multiple stories and ignore grammatical literalism in defining an event. Machine coding is more transparent because it does not do this, and therefore we think it is more reliable. Machines do not infer implied events and they do not miss events simply because they are entangled grammatically with another event. For example, a police action against protestors will not be coded as a 'protest demonstration' unless grammatically the protest is also presented in a full noun-verb clause of the form: who (source) did what (event) to whom (target). Human coders might (inconsistently) code the 'protesting students' who were the target of the police action, but the machine will not unless programmed to do so.

Automated coding entails the hazard of duplication. If the same event is reported in multiple stories, the machine will generate multiple event records. Certainly multiple

reports, with nuanced distinctions, are pervasive in virtually every event database. A common example is the 'near-duplicate', where slight changes in grammatical presentation make the components of an event distinct. At the variable level of source-event-target, there is a near equivalence of, for example, a USA-ORGA-POL (the IDEA code for 'United States', 'government agency', 'police officer') accusing a SAU-GROU-BUS ('a Saudi Arabian', 'group', 'businesses') of being a front for a terrorist ring and the 'same' general event reiterated by a USA-ORGA-EXE (i.e. a chief executive or White House spokesperson on the same day and in the same city). Slight changes in the grammatical presentation of an event may create 'near-duplicate' event records that a human coder would probably treat as a duplicate. The risk is greatest with crisis events, such as a coup d'état, or a protracted process, such as a national election, that generate repeated references to the same real world events or processes, often filed by news reporters on the same or subsequent days. Human review is the only technique that can fully identify these, but our experience is that they are concentrated in specific event forms, limiting the scope of the necessary human review.

This clause unit of analysis is an important characteristic of current machine coding technology for developing events data. With future refinement, the unit of analysis will likely shift toward a more thematic unit at the level of paragraphs or even a topic/issue unit at the level of whole documents. At this time, the analyst needs to recognize the possible importance of duplicates, given their research question, and develop a strategy of machine and human review to control for these.

The VRA[®] Knowledge Manager system works explicitly and exclusively with the material presented in the reports. It does not bring to the parsing task a repertoire of

knowledge specific to particular contexts. Indeed, we have striven to develop the IDEA protocol in a context-independent manner. Where a regional or area expert would draw upon a vast knowledge base while coding, the automated software system must rely on a much leaner set of rules and terms of reference during its parsing and coding processes. This means that nuance and context-specificity are lost. But complete consistency and transparency are gained. In reliability tests, Schrodtt & Gerner (2001) found that contextually knowledgeable human coders missed a larger share of the events than the machine, owing to fatigue, misunderstanding of grammar, and misapplication of coding rules. This parallels King & Lowe's (2003) tests of the VRA[®] Reader applied to Reuters reports of events in Bosnia. The resulting data are therefore useful for comparative analyses but not for in-depth contextual understanding.

In addition to *who does what with/to whom*, IDEA also includes indicators of *when, where, and how* the event reportedly took place, along with some report attribute information or meta-information, such as the Reuters bureau from which it originated or its byline.

Level of Analysis

The level of analysis can vary from intrapersonal (when running the system on speeches to discern operational codes, for example) to individuals to groups and organizations. Our primary approach is to identify and assess events conducted variously by individuals, groups, and organizations with major emphasis on countries and territories as recognized in the CIA's *World Factbook*. Increasingly, we are working at the first-level administrative units within countries and are in the process of fully integrating a standardized (but constantly updated) list of these entities for the world. However, we find that extracting accurate casualty, location,

and other basic event-context and attribute information below the country level of analysis is extremely difficult – and this applies to human and well as machine coding. Ultimately, there is no system requirement that fixes the analysis at any particular level; it is driven by the needs of researchers and resource constraints.

Scope of Analysis

Here we refer to the range of event forms identified in the reports. Our efforts to date have focused on social, political, environmental, and economic event forms, with much more progress evident in the social and political than economic and environmental domains. A distinctive feature of the IDEA protocol is that the more general event forms are not bound to specific actors. This contrasts with conventional international relations coding. For example, in World Event/Interaction Survey (WEIS), a ‘reduction in relations’ refers to a specific form of diplomatic (i.e. state) behavior (McClelland, 1978), but in IDEA, a reduction in routine activity refers to any reduction of routine and planned activities, including cancellations, recalls, and postponements explicitly presented as a protest against the routine, regardless of the level of the actors involved. Thus, a divorce statement in a news release constitutes an event report that is not bound to a state (or any other level of organization) actor. By pairing the actor/target with specific events, the analyst can derive the WEIS diplomatic ‘break relations’ as well as the broader set of ‘break relations’.³

³ In this way, an event output may or may not constitute an exact cross-mapping from IDEA to one of the other event frameworks. For example, just as a country closing one of its embassies maps to the IDEA event form ‘break relations’, a couple in the process of a divorce also maps to ‘break relations’. Both IDEA and WEIS frameworks include a ‘break relations’ event form but, in order to extract the WEIS equivalent of ‘break relations’ from IDEA, one must first filter by actor, in this case a state actor. A few IDEA events, especially at the terminal level, are bound to actors. An ‘armed force naval display’, for

Throughout our adaptation and extension of the WEIS framework, we have retained its focus on the political domain, while adding substantially to the realm of social conflict, particularly in terms of protest behavior. Following our early work with PANDA, we chose to build upon WEIS primarily because its nominal level of measurement does not assume a unidimensional view of conflict, from violence to cooperation. While our early PANDA work focused on the contentious and coercive but not yet violent direct action, we did much less specification of social and political conflict resolution or what might be characterized as strategies of cooperation or accommodation.⁴ Even less work has been done on categorizing the economic, environmental, and state of being (e.g. human affect and human cognition) domains, though in the spirit of the IDEA project’s goal of extensibility, we have retained large placeholder or residual categories for further differentiation.

Who/Whom

The units of analysis for the actors (source and target of an event) include individuals, groups (including ephemeral groups like crowds), organizations (including corporate entities, both public and private), and all generally recognized countries (including states and related territories, currently numbering just over 260). We use four actor variables to indicate

- (1) the normalized name of the actors identified in the text [SrcName/TgtName];
- (2) the administrative unit of the named actor [Admin];

example, need not be restricted to a military naval display, but it is highly unlikely that it will appear as something other than a military naval display. Similarly, judicial actions require some officially sanctioned institution, typically affiliated with a state, and censorship requires mass media as a target.

⁴ CAMEO (Germer et al., 2002) represents strides in this area.

- (3) the actor's role or sector [SrcSector/TgtSector];
- (4) the actor's level of social organization [SrcLevel/TgtLevel].

It may be useful to consider the sector indicator as representing a 'horizontal' cut while the level indicator serves as a 'vertical' cut within the social, economic, environmental, and political context in which the actor is identified.

The sector variable currently contains 132 values. These sectors are divided into two basic subtypes: (1) true agents, comprising 11 civilian sectors including students, labor and ethnic groups, for example, and 35 government sectors such as the national executive, the judiciary, and the police; and (2) pseudo-agents, comprising 16 intangible sectors including military hardware and typhoons, for example, and 68 tangible sectors such as polls, historical figures, and diseases. We include tangible and intangible things because, like true agents, they can function grammatically as actors. Like IDEA event forms, IDEA sectors are arrayed in a hierarchical fashion. The IDEA sector 'true agent', for example, includes government agents and civil society agents. The insurgent sector is a subset of the armed civilian group sector which, in turn, is a subset of the civil society agents, and so on.

The level of organization variable has 18 levels of differentiation. Examples include countries, cities, capitals, individuals, groups, organizations, etc. These four variables operate together to identify the actor by country, subnational unit, and sector: the output actors are presented then as Name+Admin+Sector+Level. Finally, we also retain the (non-normalized) literal name or descriptive phrase identifying the actors. Both the normalized and non-normalized lists of actors can be embedded in the events table output or linked to it in a separate table. This allows us to separate domestic or

civil from interstate events and to gauge events that cross traditional boundaries, such as protests against foreign states and state repression targeted at foreign citizens located in another country. This is invaluable in evaluating the globalization of contentious and other politics.

The IDEA sectors also serve to organize the supplemental noun classes used in the coding process. Noun classes refer to the synonymy or the semantic relations between word forms. These relations can take the form of hyponyms (e.g. English bulldog is a hyponym [subordinate] of dog) or hypernyms (e.g. dog is the hypernym [superordinate] of English bulldog). Using WordNet's 25 unique beginners⁵ as a base, we assembled a comprehensive hierarchical listing of semantic classes arrayed in a lattice, from which the parser utilizes the grammatical 'parents' and 'children'. Rather than associate a source as a literal word or phrase (e.g. US warplanes) with a verb and target (e.g. US warplanes bombed Iraq), we simply utilize noun classes. For example, military hardware or <MILH> bombed true agent or <TAGE>. In this case, 'military hardware' contains hundreds of entries like F18, F-16, fighter jet, Blackhawk helicopters, MiG jets, tank buster aircraft, etc. Similarly, the noun class 'true agent' contains tens of thousands of entries ranging from official country names (e.g. the United States of America, US, U.S., USA, etc.) to titles (e.g. President, president, Prime Minister, PM, Mr., Dr., etc.) and other labels (e.g. prostitutes, farmers, entrepreneurs, drug dealers, prisoners, steel workers, etc.). Currently our sense index contains some 187,000 open class English words (i.e. nouns, verbs, adjectives, and adverbs).

⁵ Each of the 25 unique beginners in WordNet corresponds to 'relatively distinct semantic fields, each with its own vocabulary' (Miller, 1998: 28). Examples of unique beginners for noun source files include things like food and locations. See the WordNet website for details: <http://www.cogsci.princeton.edu/~wn/>.

Certain event forms – an apology for example – are rarely presented in their verb form. Unless the text is in the first person, one generally reads about an apology (in its noun form) issued by one party to another rather than reading that an actor apologizes (in its verb form) to another, except in the case of a direct quote included in a news report. We have integrated approximately 150 of these sector/noun classes into the IDEA protocol.⁶ This part of the protocol changes quite often as classes are added and/or modified (especially at the lower levels) to yield more detail in a specific domain, or to better deal with a particular kind of event or phenomenon.

What

The core focus of analysis for the social, political, and economic events that we code is the nominally scaled forms of behaviors in which we have an interest. Since the IDEA protocol explicitly builds upon the WEIS framework, we have retained its 22 top-level ‘cue’ categories. These ‘cue’ categories are still used by the vast majority of analysts who work with events data. As noted above, we try not to differentiate among event forms done by different actors or having particular targets, at least *a priori*. Such actor/target-specific event listings can readily be produced from a sorted output of coded events. The acronym for the IDEA events variable coded by the Reader is [EventForm]. As with the actors, the Reader also retains and can output the actual verb phrases from which the codes were derived.

Descriptions, examples, and usage notes for each of the roughly 250 current IDEA event forms can be found at <http://vranet.com/IDEA>. About 150 IDEA events are considered terminal; that is, at the

current level of automated coding technology, no further detail can be differentiated.⁷

When

The date that the event occurred is assumed to be the date of the report, unless specified otherwise in the text. Thus, most of the event date codes come directly from the report date. However, when a modifying phrase such as ‘last week’s riot’ or ‘the meeting next week’ is found, the event date recorded by the parser will diverge from the report date by simply subtracting or adding as appropriate from the date of the report. The variable indicating when the event happened is simply called [Date]. We are currently programming the Reader to distinguish current from future or past, based on verb tense, so it should be possible in the future to distinguish past events from future events.

Where

The precise location of an event is extremely difficult to identify in many news report leads, both for humans and for machines. More often than not, no explicit reference to location is carried in the first lines of a report. Rather, this information is most often embedded in the header of the report, particularly the headline, bureau, dateline, or byline. In addition, it is sometimes buried deep within a more lengthy report, often by a reference to another actor and/or event; for example, the location information is implicitly conveyed by reference to specific actors in ‘Iraq invaded Kuwait’. This indirect means of referencing location is sufficient for many, but not all, analyses.

In sum, we make a systematic attempt to identify the specific place of the events from the leads. Most often, the system finds it in

⁶ A complete listing of sectors and levels of organization along with their descriptions can be found at <http://vranet.com/idea/coderhelp/testcoderhelp.htm> under the heading variables.

⁷ ‘Biological weapons use’ and ‘chemical weapons use’ are both examples of terminal events. Finer gradations are not currently provided. Thus, an anthrax attack would map to ‘biological weapons use’.

the location associated with an actor or the header information. Less often (>20%), there is a prepositional phrase marking the place. The location variable is called [Place]. At present, the system outputs about 270 standardized names of countries and related territories. We are experimenting with various standards for outputting first level administrative unit information, and we currently use a combination of the National Imagery and Mapping Agency (NIMA) and the CIA's *World Factbook* codes.

Reliability

VRA's last formal reliability in-house test was conducted in September 2000. The results ranged from 70% to 80%, depending on the basis for comparison. These results are comparable (indeed favorable if one considers the type of error) with large-scale human coding efforts. In an independent test, King & Lowe (2003) obtained comparable reliability from the Reader to human coders. These ranged from 60% to 80%, with higher reliability at the 'cue' level. We have also tracked progressive improvements in coding reliability over time. In a more recent test of events involving use of force in Egypt and Tajikistan, Jenkins, Abbott & Taylor (2002) find terminal level reliabilities in the 80–90% range.

The major advantage of automated coding is speed. According to Gerner et al. (2002: 2), 'human coders typically produce between 5 and 10 events per hour'. A dense dataset like India, for example, contains upwards of 194,554 events between 1 January 1990 and 1 July 2002. Assuming a typical human coder can code 7.5 events per hour, it would take approximately 12.5 years for one coder working 40 hours per week to code India, whereas the parser can code the same dataset in less than a day. A human coding endeavor of this magnitude would require immense oversight in terms of coder training and quality control, not to mention

financial outlay. It also disregards the reality of coder fatigue and the possibility of rogue coders, both of which can significantly diminish the overall reliability of the data.

A key advantage of automation is that protocol improvements are likely to be permanent and cumulative. This does not mean that the progress is steady. It can be reversed if changes alter the coding of other events and are not fully tested before use. This type of context-free coding will inevitably entail some error, but our experience and that of others is that it is better than the normal error of human coding. We have developed an extensive system of supplemental noun classes to leverage our ongoing protocol development efforts. In a recent case, around five hundred additional events were identified in one country-year after the introduction of a single verb complement frame. The added frame represented a very common syntactic and semantic pattern in the particular set of reports.⁸

Future Developments in Event Data

The IDEA conceptual framework offers a useful extension of a human events coding tradition that extends back nearly 40 years. We have sought, throughout our development process, to preserve backwards compatibility as well as extensibility. We have built upon the nominally scaled WEIS framework because we think the constraint of fitting events into a one-dimensional conflict-cooperation array such as COPDAB is ill-advised. It seems better to reduce the number of assumptions built into an event framework and focus on getting the events 'right' in terms of conceptual clarity. By developing events data spanning the full

⁸ King & Lowe (2003), in an independent test of the VRA[®] Reader's coding performance, found that automated coding was as accurate as trained coders, but they argued that the machine would be far better in the long run, owing to the difficulty of finding and training qualified coders who could stay with the job over the long haul.

breadth of social, economic, and political event forms, and including event attributes that tap into the rich multidimensionality of violent and nonviolent, contentious and routine, and coercive and accommodating behaviors, we hope to build upon the best in the events data tradition.⁹ The four-level event hierarchy of IDEA provides flexibility as well as conceptual and coding clarity. IDEA also includes a variety of dimensions, such as the contentiousness and coerciveness of events and other event attributes, which can be used to measure international interactions.

Finally, the 'garbage in garbage out' adage must be acknowledged. As noted above, our unit of analysis is the *clause-bound event report*. One must weigh the report sources against one's research or other interests. Certainly multiple input sources of text should be considered, especially local-language news reports. Several researchers have found that local and international news sources provide different pictures of political processes that ultimately need to be combined (Sommers & Scarritt, 1999; Schrodt & Gerner, 2001; Davenport & Ball, 2002). At this point in automated coding, it is not clear how multiple sources can be integrated, except at the gross comparison level. We suggest a representative selection of international, national, and local media will be needed to improve our understanding of the actual situations on the ground. We also suggest supplementing the tracking of news reports with the tracking of field situation and incident reports. In this way, we can begin to triangulate the observations from above (international news agencies) with those from below (by personnel actually working in the field) to better track conflict situations that may erupt into violence and support

means to intervene earlier to mitigate the destructive consequences.

References

- Azar, Edward E., 1980. 'The Conflict and Peace Data Bank (COPDAB) Project', *Journal of Conflict Resolution* 24(1): 143–152.
- Bond, Doug & Bill Vogeley, 1995. 'Profiles of International Hotspots'. Unpublished manuscript. Center for International Affairs, Harvard University.
- Bond, Doug; J. Craig Jenkins, Charles Lewis Taylor & Kurt Schock, 1997. 'Mapping Mass Political Conflict and Civil Society: Issues and Prospects for Automated Development of Event Data', *Journal of Conflict Resolution* 41(4): 553–579.
- Bond, Joe & Doug Bond, 1995. *Panda Codebook*. Cambridge, MA: The Program on Nonviolent Sanctions and Cultural Survival, Weatherhead Center for International Affairs, Harvard University.
- Davenport, Christian Alexander & Patrick Ball, 2002. 'Implications of Source Selection in the Case of Guatemalan State Terror', *Journal of Conflict Resolution* 46(3): 427–450.
- Eriksson, Mikael; Peter Wallensteen & Margareta Sollenberg, 2003. 'Armed Conflict, 1989–2002', *Journal of Peace Research* 40(5): 593–607.
- Esty, Daniel C.; Jack A. Goldstone, Ted Robert Gurr, Barbara Harff, Marc Levy, Geoffrey D. Dabelko & Pamela Surko, 1998. *State Failure Task Force Report: Phase II Findings*. McLean, VA: Science Applications International Corporation.
- Gerner, Deborah J.; Philip A. Schrodt, Ronald A. Francisco & Judith L. Weddle, 1994. 'The Machine Coding of Events from Regional and International Sources', *International Studies Quarterly* 38(1): 91–119.
- Gerner, Deborah J.; Philip A. Schrodt, Rajaa Abu-Jabr & Ömür Yilmaz, 2002. 'Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions', paper presented at the 43rd Annual Convention of the International Studies

⁹ For a complete listing of the output variables, including the five event attributes of the domain of action, affect, mechanism of action, physical injury, and damage, see <http://www.vranet.com/idea/output>.

- Association, New Orleans, LA, 24–27 March.
- Goldstein, Joshua S., 1992. 'A Conflict–Cooperation Scale for WEIS Events Data', *Journal of Conflict Resolution* 36(3): 369–385.
- Goldstein, Joshua & Jon C. Pevehouse, 1996. 'Reciprocity, Bullying and International Cooperation: A Time-Series Analysis of the Bosnia Conflict', *American Political Science Review* 91(3): 515–530.
- Henderson, Conway W., 1991. 'Conditions Affecting the Use of Political Repression', *Journal of Conflict Resolution* 35(1): 120–142.
- Jenkins, J. Craig & Doug Bond, 2001. 'Conflict Carrying Capacity, Political Crisis and Reconstruction: A Framework for the Early Warning of Political System Vulnerability', *Journal of Conflict Resolution* 45 (February): 3–31.
- Jenkins, J. Craig; Marianne Abbott & Charles L. Taylor, 2002. 'Constructing International Conflict Indicators: A Reliability Assessment of Automated Coding for *The World Handbook of Political and Social Indicators IV*', paper presented at the 43rd Annual Convention of the International Studies Association, New Orleans, LA, 24–27 March.
- Jones, Daniel M.; Stuart A. Bremer & J. David Singer, 1996. 'Militarized Interstate Disputes, 1816–1992: Rationale, Coding Rules, and Empirical Patterns', *Conflict Management and Peace Science* 15(2): 163–213.
- King, Gary & Will Lowe, 2003. 'An Automated Information Extraction Tool For International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design', *International Organization* 57(3): 617–642.
- McClelland, Charles A., 1978. 'World Event/Interaction Survey (WEIS) Project, 1966–1978', Third ICPSR Edition. Ann Arbor, MI: Inter-University Consortium for Political and Social Research.
- McClelland, Charles A., 1983. 'Let the User Beware', *International Studies Quarterly* 27(2): 169–177.
- Miller, George A., 1998. 'Nouns in WordNet', in Christiane Fellbaum, ed., *Wordnet: An Electronic Lexical Database*. Cambridge, MA: MIT Press (23–46).
- Poe, Steven G. & C. Neal Tate, 1994. 'Repression of Human Rights to Personal Integrity in the 1980s: A Global Analysis', *American Political Science Review* 88(4): 853–872.
- Russett, Bruce; Hayward R. Alker, Karl W. Deutsch & Harold D. Lasswell, 1964. *World Handbook of Political and Social Indicators*. New Haven, CT: Yale University Press.
- Schrodt, Philip A. & Deborah Gerner, 1994. 'Validity Assessment of a Machine-Coded Event Data Set for the Middle East, 1982–92', *American Journal of Political Science* 38(3): 825–854.
- Schrodt, Philip A. & Deborah Gerner, 2000. 'Cluster-Based Early Warning Indicators for Political Change in the Contemporary Levant', *American Political Science Review* 94(4): 803–818.
- Schrodt, Philip & Deborah Gerner, 2001. 'Automated Coding of International Event Data Using Sparse Parsing Techniques', paper presented at the 42nd Annual Convention of the International Studies Association, Chicago, IL, 20–24 February (<http://www.ukans.edu/~keds/pdf.dir/TABARI.ISA01.pdf>).
- Sommers, Henrik & James R. Scarritt, 1999. 'The Utility of Reuters for Events Analysis in Area Studies: The Case of Zambia–Zimbabwe Interactions, 1982–1993', *International Interactions* 25 (Spring): 1–31.
- Taylor, Charles Lewis & Michael C. Hudson, 1972. *World Handbook of Political and Social Indicators: Second Edition*. New Haven, CT: Yale University Press.
- Taylor, Charles Lewis & David A. Jodice, 1983. *World Handbook of Political and Social Indicators: Third Edition*. New Haven, CT: Yale University Press.
- DOUG BOND, b. 1954, PhD in Political Science (University of Hawaii, 1985); Associate Director, Program on Nonviolent Sanctions and Cultural Survival; Lecturer in Extension, Harvard University Division of Continuing Education; President, Virtual Research Associates, Inc. Current main interests: automated events data development, Korea.

JOE BOND, b. 1963, PhD in Political Science (Purdue University, 1994); Affiliate, Program on Nonviolent Sanctions and Cultural Survival; Instructor, Harvard University Division of Continuing Education; Vice President, Virtual Research Associates, Inc. Current main interests: political psychology, data mining and automated events data development, Balkan politics.

CHURL OH, b. 1961, PhD in Chemistry (Boston University, 1994); Affiliate, Program on Nonviolent Sanctions and Cultural Survival; Vice President of Software Development, Virtual Research Associates, Inc. Current main interest: software development designed to draw, model, and manage chemical structure data, and to integrate these applications into the World Wide Web environment.

J. CRAIG JENKINS, b. 1948, PhD in Sociology (State University of New York, Stony Brook, 1975); Full Professor, Department of Sociology and Political Science (by courtesy), Faculty Associate, Mershon Center for International Security, and Research Fellow, Center for African Studies, Ohio State University. Most recent books: *The Politics of Insurgency: The Farm Worker Movement of the 1960's* (1985); *The Politics of Social Protest* (1995).

CHARLES LEWIS TAYLOR, b. 1935, PhD in International Relations (Yale University, 1963); Full Professor, Department of Political Science, Virginia Polytechnic Institution and State University; co-author of the second and third editions of the *World Handbook of Political and Social Indicators* (1972, 1983). Current main interest: cross-national analysis of political conflict.