



ORIGINAL ARTICLE

Introducing ICBe: an event extraction dataset from narratives about international crises

Rex W. Douglass¹, Thomas Leo Scherer¹ , J. Andrés Gannon² , Erik Gartzke¹, Jon Lindsay³, Shannon Carcelli⁴, Jonathan Wilkenfeld⁴, David M. Quinn⁴, Catherine Aiken⁵, Jose Miguel Cabezas Navarro⁶, Neil Lund⁴, Egle Murauskaite⁴ and Diana Partridge⁴

¹Department of Political Science, University of California, San Diego, CA, USA, ²Department of Political Science, Vanderbilt University, Nashville, TN, USA, ³School of Cybersecurity and Privacy, Georgia Institute of Technology, Atlanta, GA, USA, ⁴Department of Government and Politics, University of Maryland, College Park, MD, USA, ⁵Center for Security and Emerging Technology, Georgetown University, Washington, DC, USA and ⁶Health and Society Research Center, Universidad Mayor, Santiago, Chile

Corresponding author: Rex W. Douglass; Email: rexdouglass@gmail.com

(Received 6 November 2022; revised 29 November 2023; accepted 9 February 2024; first published online 24 May 2024)

Abstract

How do international crises unfold? We conceptualize international relations as a strategic chess game between adversaries and develop a systematic way to measure pieces, moves, and gambits accurately and consistently over a hundred years of history. We introduce a new ontology and dataset of international events called ICBe based on a very high-quality corpus of narratives from the International Crisis Behavior (ICB) Project. We demonstrate that ICBe has higher coverage, recall, and precision than existing state of the art datasets and conduct two detailed case studies of the Cuban Missile Crisis (1962) and the Crimea-Donbas Crisis (2014). We further introduce two new event visualizations (event iconography and crisis maps), an automated benchmark for measuring event recall using natural language processing (synthetic narratives), and an ontology reconstruction task for objectively measuring event precision. We make the data, supplementary appendix, replication material, and visualizations of every historical episode available at a companion website crisisevents.org.

Keywords: data collection; measurement; text and content analysis

If we could record every international interaction in the realms of diplomacy, conflict, economics, and beyond, how much unique information would this chronicle amount to, and how surprised would we be to see something new? In other words, what is the entropy of international relations? While this record could in principle be unbounded, the central conceit of social science is that there are structural regularities that limit what actors can do, their best options, and even which actors are likely to survive (Brecher, 1999; Reiter, 2015). If so, then these events can be recorded and systematically measured by social scientists interested in these regularities.¹ A large and growing measurement literature seeks to do just that, using human coding and

¹See work on crises (Brecher and Wilkenfeld, 1982; Beardsley *et al.*, 2020), militarized disputes (Gibler, 2018; Palmer *et al.*, 2022), wars (Reiter *et al.*, 2016), organized violence (Sundberg and Croicu, 2016; Davies *et al.*, 2022), political violence (Raleigh *et al.*, 2010), sanctions (Felbermayr *et al.*, 2020), and international agreements (Kinne, 2020; Owsiak *et al.*, 2018), dispute resolution (Frederick *et al.*, 2017), and diplomacy (Moyer *et al.*, 2021; Sechser, 2011).

improving natural language processing techniques to capture unstructured streams of events from text such as international news reports.²

We advance existing efforts to identify and structure regularized events and actors in international politics by combining human coding with natural language processing to create (1) a large, flexible ontology of international affairs and (2) a fine-grained and structured event dataset of international crises from 1918 to 2017, which we developed by applying our ontology to an unusually high-quality corpus of historical narratives of international crises (Brecher, 1999; Wilkenfeld and Brecher, 2000; Brecher *et al.*, 2016). We then develop several methods for objectively gauging how well these event codings reconstruct the information contained in the original crisis narrative. We conclude by benchmarking our event codings against several current state-of-the-art event data collection efforts. The underlying fine-grained variation in international affairs is unrecognizable through the lens of current quantification efforts. We find that existing models produce data on historical episodes that do not contain enough information to reconstruct the underlying event. In focusing this initial effort on international crises as a proof of concept sample, we demonstrate our ontology and method's potential to improve upon existing empirical identifications of patterns of international interactions.

Over the next five sections, this measurement paper makes the following arguments. First, there is a real-world unobserved latent concept known as international relations that can and should be systematically measured. Second, we propose a method for systematic large-scale measurement of the actors and behaviors in international affairs and as a proof of concept apply that method to a well-regarded and salient sample of events known as international crises. Third, in doing so, we confirm that those measurements exhibit several desirable kinds of internal and external validity and out-perform existing approaches. Fourth, this validation can be evaluated in detail via new event visualizations, with examples provided for case studies of the 1962 Cuban Missile Crisis and 2014 Crimea-Donbas Crisis. A final section concludes.

1. Identifying and measuring international relations

1.1 Motivation

Our knowledge of any historical episode, including the participants and their preferences, behaviors, and beliefs, is only indirectly observed from historical records that most often take the form of unstructured natural language text. Despite its complexity, all international interactions fundamentally involve a finite set of actors expressing their interests through at least theoretically observable behaviors. So how can we abstract and measure discrete events that make up a historical episode in international relations? The easiest way to convey the desired product is with an example. [Figure 1](#) shows a narrative account of the Cuban Missile Crisis (1962) in natural language sentences alongside a mapping to discrete machine-readable abstractive events. From this, scholars can identify similarities and differences across events like what foreign policy actions deter versus inflame (Jervis, 1978; Glaser, 2000), when third parties mediate (Haffar, 2002; Quinn *et al.*, 2006), and how actors communicate resolve (Trager, 2016; Lupton, 2018). Identifying patterns of international interactions is not just an inherently interesting enterprise; it is a necessary precondition to important efforts to predict where policymakers should turn their attention to improve global welfare (Ward *et al.*, 2013; Beger *et al.*, 2021).

1.2 Existing state of the art measurements

We begin by drawing informative prior beliefs about the underlying process of international relations that we expect to govern behavior during historical episodes and their later transcription into the historical record. We organize our prior beliefs along two overarching

²See Beiler *et al.* (2016), Boschee *et al.* (2015), Brandt *et al.* (2018), Grant *et al.* (2017), Li *et al.* (2021). On event extraction from images and social media see Zhang and Pan (2019) and Steinert-Threlkeld (2019).

S	Natural Language Sentences (ICB Corpus)	Machine-Readable Events (ICBe)
4	When the U.S. discovered the presence of Soviet military personnel in Cuba on 7 September 1962 it called up 150,000 reservists.	mobilization 100ks
5	The Soviets mobilized on the 11th.	discover fact - deployment to area 100s:1ks
8	The U.S. crisis was triggered on 16 October when the CIA presented to President Kennedy photographic evidence of the presence of Soviet missiles in Cuba.	mobilization 1ks discover fact:start of crisis - deployment to area fortify
9	The U.S. responded with a decision on the 20th to blockade all offensive military equipment en route to Cuba.	start of crisis - deployment to area blockade coastline 10ks
10	When this was announced on 22 October, a crisis was triggered for Cuba and the USSR. An urgent meeting of the UN Security Council was requested by both the U.S. and Cuba on the 22nd, and by the USSR the next day.	blockade coastline start of crisis
11	On the 23rd as well, the Soviets accused the United States of violating the UN Charter and announced an alert of its armed forces and those of the Warsaw Pact members.	appeal meeting raise in alert 100ks accuse violate terms of treaty
12	That day Cuba responded by condemning the U.S. blockade and declaring its willingness to fight.	disapprove coastline
13	A resolution was adopted on the 23rd by the OAS calling for the withdrawal of the missiles from Cuba and recommending that member-states take all measures, including the use of force, to ensure that the government of Cuba would not continue to receive military material.	end military cooperation
14	On 24 October the Security Council adopted a resolution requesting the Secretary-General to confer with the parties.	demand withdraw from area 1ks express intent mediation
15	On that same day, U Thant began mediation by sending identical letters to Khrushchev and Kennedy which proposed that the Soviet Union and the United States enter into negotiations, during which period both the shipment of arms and the quarantine would be suspended.	express intent - mediation
16	Moscow's major response to the crisis was a letter from Khrushchev to Kennedy on 26 October offering the removal of Soviet offensive weapons from Cuba and the cessation of further shipments in exchange for an end to the U.S. quarantine and a U.S. assurance that it would not invade Cuba.	offer happens - withdraw from area -C- will happen end blockade
17	The situation was exacerbated on the 27th when a U.S. U-2 surveillance plane was shot down.	offer happens - withdraw from area -C- will happen end blockade border violation airspace 1s 1ks
18	That day another Khrushchev letter was received in Washington offering the removal of Soviet missiles from Cuba in exchange for the removal of U.S. missiles from Turkey.	offer happens - withdraw from area -C- will happen withdraw from area
19	U.S. mobilization and aerial reconnaissance flights were stepped up.	offer happens - withdraw from area -C- will happen border violation:mobilization mobilization airspace
20	And on the 27th President Kennedy sent the Soviet premier an acceptance of the proposals contained in the letter of 26 October while making no reference to Khrushchev's second letter of the 27th.	accept
21	The following day Khrushchev notified the U.S. government that he had ordered work on the missile sites in Cuba stopped.	accept
23	At the same time he warned Washington that U-2 reconnaissance flights over Cuba must be stopped as well.	demand withdraw from area airspace
24	The crisis continued at a lower level of intensity for several more weeks due to Cuban President Castro's demands concerning a U.S. pledge not to invade his country.	demand de-mobilization:lower alert body of water:coastline 100ks
25	On 30 October U Thant began talks in Havana, and Kennedy agreed to lift the quarantine for the duration of the talks.	end blockade coastline 10ks
26	When Cuba rejected UN inspection, the U.S. resumed the quarantine and air surveillance.	accept end blockade coastline reject unspecified cooperation
27	The Kremlin sent Deputy Premier Anastas Mikoyan to Cuba on 2 November to try to persuade Castro to allow UN inspection.	blockade airspace:coastline 10ks
28	When this proved unsuccessful, a U.S.-USSR agreement was reached on 7 November allowing U.S. inspection and interception of Soviet ships leaving Cuba and the photographing of the missiles.	discussion:meeting
29	The following day the superpowers negotiated the removal of the IL-28 bombers which Castro had claimed were Cuban property.	withdraw from area 100s
30	Castro's agreement was conveyed to the U.S. on 20 November 1962, which terminated the Missile crisis for all three actors.	sign formal agreement end of crisis
31	The U.S. naval quarantine was lifted immediately, but aerial surveillance continued until the agreement was completely carried out.	end blockade body of water:coastline 10ks

Figure 1. Comparison of a natural language and machine-readable abstractive account of the Cuban Missile Crisis (1962). The text on the left is a summary of the event from the ICB Crisis Narrative. The mapping on the right shows the corresponding ICBe coding.

Downloaded from https://www.cambridge.org/core. IP address: 3.141.30.183, on 12 Mar 2025 at 15:39:14, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/psrm.2024.17

axes: (1) existing efforts to identify the actors/actions of international relations; and (2) the types of behaviors and information we hope to recover. [Table 1](#) describes these two axes as columns and rows, respectively.

The rows in [Table 1](#) represent the types of information we expect to find in international relations and forms the basis for our proposed ontology. We began the ontology by first doing a full natural language processing pass of the corpus and identifying all of the named entities and verbs mentioned in the text. To identify possible behaviors, we matched verbs to the most likely definition found in Wordnet (Miller, 1995), tallied them (SI Appendix 1.2), and then aggregated them into a smaller number of behaviors balancing conceptual detail with manageable sparsity for human coding (informed by existing conceptual literature and measurement research). We used the International Crisis Behavior (ICB) project actor level data to identify likely actors for each crisis and location options relative to each actor. For behavior, actor, and location, coders could write-in a value if the given options were insufficient. The codebook lists eleven behaviors added post-coding as coders flagged events that were not captured by the initial ontology (e.g., propaganda).

As we are not the first to attempt to measure international relations in a structured manner, the columns of [Table 1](#) compare the ontological coverage of ICBe to existing state of the art systems in production and with global coverage. We choose these datasets and models as they represent frequently used and reputable efforts to structure and describe historical events of interest to scholars of international politics. The first column starts with our contribution, ICBe, alongside other event-level datasets including CAMEO dictionary lookup-based systems (Historical Phoenix (Althaus *et al.*, 2019); ICEWS (Boschee *et al.*, 2015); Terrier (Grant *et al.*, 2017)), the Militarized Interstate Disputes Incidents dataset, the UCDP-GED dataset (Sundberg and Melander., 2013; Davies *et al.*, 2022), and ACLED (Raleigh *et al.*, 2010).³ The final set of columns compares episode-level datasets beginning with the original ICB project (Brecher *et al.*, 2016; Brecher and Wilkenfeld, 1982; Beardsley *et al.*, 2020), the Militarized Interstate Disputes dataset (Gibler, 2018; Palmer *et al.*, 2022), and the Correlates of War (Sarkees and Wayman, 2010). We include episode-level datasets as they remain a common and trusted tool for analyzing international relations, and because ICBe is unique among event-level datasets as events are matched to crises and can be aggregated to the episode level. There is imperfect overlap between their intended depth and scope of coverage; “international crises” are similar, but not identical to, “interstate wars” and “militarized interstate disputes,” which differ yet again from “individual events of organized violence” and “non-violent action.” Even like-concepts require care in comparison, as an “aim” in ICBe is the same as in MIPS, but an “alert” in ICBe is not the same as an “alert” in MIDs.

This comparison is not intended to fault existing data and models for not including every variable in ICBe’s ontology, as some of these variables fall outside the scope of a particular dataset’s intended purpose. Rather, it serves as an initial basis for identifying the heterogeneity in existing efforts to abstract and measure discrete historical events of interest and to provide theoretical justifications from existing research about what is included in our dataset’s ontology and where ICBe’s detail about historical events can be compared to the current state of the art.

With the exception of large-scale CAMEO dictionary-based systems (the first grouping of columns), our ontology improves upon the existing state of the art quantitative datasets that ignore important information about international interactions.⁴ We highlight two particular innovations. First, we separate the “chess pieces” from the “chess players” in distinguishing between different actors within a state. By virtue of our ontology, coding military versus civilian

³Other related datasets that insufficiently overlap ICBe’s domain for comparison include BCOW (Leng and Singer, 1988), WEIS (McClelland, 1978), CREON (Hermann, 1984), CASCON (Bloomfield and Moulton., 1989), SHERFACS (Sherman, 2000), Real-Time Phoenix (Brandt *et al.*, 2018), and COFEE (Balali *et al.*, 2021) (see histories in Merritt, 1994 and Schrodtt and Hall, 2006).

⁴See Balali *et al.* (2021) for a recent review of ontological depth and availability of Gold Standard example text.

Table 1. Ontological coverage of ICBe versus the existing state of the art

Concept		Literature	ICBe	Phoenix	Terrier	ICEWs	MID					
							incidents	UCDP-GED	ACLED	ICB	MIDs	COW
Domain	Type (episode or event)		Ev	Ev	Ev	Ev	Ev	Ev	Ev	Ep	Ep	Ep
	Start		1918	1945	1977	1995	1993	2015	1997	1918	1816	1816
	End		2017	2019	2018	2020	2010	2022	2023	2017	2014	2007
	N		32K	8.5M	28.4M	17.5M	9.6K	128K	1M	1K	5.9K	1K
	Coders (hand or automated)		H	A	A	A	H	H	H	H	H	H
	Corpus		ICB	News	News	News	Mix	News	Mix	Mix	Mix	Mix
Players	Date source (event or article)		E	A	A	A	E	A	E	E	E	E
	Location source (event or actor)		E	E	E	E	A	E	E	A	E	A
	States	Fazal (2011), Spruyt (1996)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Subnational actors	Haffar (2002)	✓	✓	✓	✓		✓	✓			✓
	IGO/NGO	Bush and Hadden (2019)	✓	✓	✓	✓			✓	✓		✓
	Civilians	Ben-Yehuda and mishali-ram (2006)	✓	✓	✓	✓		✓	✓			
	Fatalities	Lacina (2006), McNabb Cochran and Long (2017)	✓				✓	✓	✓	✓	✓	✓
	Force size	Carafano (2014), Goertz and Diehl (1986)	✓									
	Force domain	Gartzke and Lindsay (2019), Lindsay and Gartzke (2020)	✓	✓	✓	✓						
	Geography (location, territorial change)	Carter (2010)	✓						✓			
Think	Alert (start/end crisis)	Brecher (1999)	✓								✓	
	Wishes (desire/fear)	Goldgeier and Tetlock (2001)	✓								✓	
	Evaluation (victory/defeat)	Stein and Russett (1980)	✓								✓	
	Aims (territory, policy, regime, preemption)	Sullivan (2007)										
	Awareness (discover, become convinced)	Ramsay (2017), Yarhi-Milo (2013)	✓									
Say	React to past event (praise, disapprove, accept, reject, accuse)	O'Neill (2018)	✓	✓	✓	✓					✓	
	Request future event (appeal, demand)	Zartman and Olivier Faure (2005)	✓	✓	✓	✓	✓					
	Predict future event (promise, threaten, express intent, offer without condition)	Sechser (2011)	✓	✓	✓	✓	✓					✓
	Predict with condition (offer, ultimatum)	Powell (2002)	✓									

(Continued)

Table 1. (Continued.)

			MID									
Concept		Literature	ICBe	Phoenix	Terrier	ICEWs	incidents	UCDP-GED	ACLED	ICB	MIDs	COW
Do-unarmed	Government (leadership/institution change, coup, assassination)	Goemans <i>et al.</i> (2009)	✓	✓	✓	✓			✓			
	By civilians (protest/riot/strike)	Chenoweth <i>et al.</i> (2019)	✓	✓	✓	✓			✓			
	Against civilians (terrorism, domestic rights, mass killing, evacuate)	Eck and Hultman (2007), LaFree and Dugan (2007)	✓	✓	✓	✓			✓			
	Diplomacy (discussion, meeting, mediation, break off negotiations, withdraw/expel diplomats, propoganda)	Beardsley (2011)	✓	✓	✓	✓				✓		
	Legal agreements (sign agreement, settle dispute, join war on behalf of, ally, mutual defense pact, open border, cede territory, allow inspections, political succession, leave alliance, terminate treaty)	Gibler and Sarkees (2004), Owsiak <i>et al.</i> (2018)	✓	✓	✓	✓			✓	✓	✓	
	Violate agreement (violate terms of agreement)	Leeds (1999)	✓									
	Mutual cooperation (economic cooperation or aid, military cooperation, intelligence cooperation, unspecified)	Leeds (1999), Yarhi-Milo <i>et al.</i> (2016)	✓	✓	✓	✓						
Do - armed	Directed aid (general political support, economic aid, humanitarian aid, military aid, intelligence aid, unspecified aid)	Yarhi-Milo (2013)	✓	✓	✓	✓				✓		
	Preparation (alert, mobilization, fortify, exercise, weapons test)	Lai (2004)	✓	✓	✓	✓	✓		✓		✓	
	Maneuver (deployment, show of force, blockade, no fly zone, border violation)	Allen <i>et al.</i> (2022)	✓	✓	✓	✓	✓		✓		✓	
	Combat (battle/clash, attack, invasion/occupation, bombard, cease fire, retreat)	Fortna (2018), Min (2021)	✓	✓	✓	✓	✓		✓	✓	✓	✓
	Strategic (declare war, join war, continue fighting, surrender, end war, withdraw from war, switch sides)	Sarkees and Wayman (2010), Reiter (2015)	✓	✓	✓	✓	✓				✓	
Autonomy (assert political control over, assert autonomy against, annex, reduce control over, decolonize)	Frederick <i>et al.</i> (2017)	✓	✓	✓	✓	✓		✓				

actors and national leaders versus bureaucrats, our data can be used to explore important questions concerning civilian-military relations (Narang and Talmadge, 2018), Track Two diplomacy, the role of sub-national actors (Hsu *et al.*, 2020), and the evolution of which actors are engaged in crises—a topic of increasing interest as states engage in gray zone conflict by employing the coast guard or paramilitary mercenaries instead of internationally recognized state militaries (Gannon, 2022). Second, we add information about the domains in which actors behave—whether in land, air, sea, space, or cyber—since they differ in their technology, tactics, geography, and purpose (Gartzke and Lindsay, 2019). Doing so allows researchers to identify and explain patterns in escalation conditional on the military means states use in conflict. Recent concerns about cross-domain conflict, and the effect of new domains of conflict like space and cyber, have made this an endeavor of increased interest to practitioners (Gannon, 2022).

2. Methodology and data

2.1 Corpus

For our corpus, we select a set of unusually high-quality historical narratives from the ICB project ($n = 471$) with coverage spanning 1918–2017 (SI Appendix 1.1) (Brecher and Wilkenfeld, 1997; Brecher *et al.*, 2016). ICB defines a crisis as meeting three conditions: (1) an actor perceives a threat to one of more of its core values, (2) the actor has a finite time horizon for responding to the perceived threat, and (3) the probability of military hostility has increased (Brecher and Wilkenfeld, 1982). Crises are a significant focus of detailed single case studies and case comparisons because they provide an opportunity to examine behaviors in international relations short of, or at least prior to, full conflict (Holsti, 1965; Paige, 1968; Allison and Zelikow, 1971; Brecher and Wilkenfeld, 1982; Gavin, 2014; Iakhnis and James, 2019). The corpus is also unique in that it was designed to be used in a downstream quantitative coding project, meaning each narrative was written by a small number of scholars using a uniform coding scheme where things like word choice, writing style, and level of specificity were done deliberately and consistently (Hewitt, 2001). Case selection was exhaustive based on a survey of world news archives and region experts, cross-checked against other databases of war and conflict, and non-English sources (Brecher *et al.*, 2016; Kang and Yu-Ting Lin., 2019, 59).

2.2 Coding process

The ICBe ontology follows a hierarchical design philosophy where a smaller number of significant decisions are made early on and then progressively refined into more specific details (Brust and Denzler, 2020).⁵ Each coder was instructed to first thoroughly read the full crisis narrative and then presented with a custom graphical user interface (GUI) (SI Appendix 2.1). Coders then proceeded sentence by sentence, choosing the number of events (0–3) that occurred, the highest behavior (thought, speech, or action), a set of players, whether the means were primarily armed or unarmed, whether there was an increase or decrease in aggression (uncooperative/escalating or cooperative/de-escalating), and finally one or more specific and non-mutually exclusive activities. Some additional details were always collected (e.g., location and timing) while other details were only collected if appropriate (e.g., force size, fatalities, domains, units). While each event was matched to a sentence, coders could fill in details outside that sentence (e.g., antecedents to pronouns). We reviewed, standardized, and normalized where coders listed a behavior, actor, or location outside the ontology.⁶

A unique feature of the ontology is that thought, speech, and do behaviors can be nested into combinations, e.g. an offer for the U.S.S.R. to remove missiles from Cuba in exchange for the U.S.

⁵This process quickly focuses the coder on a smaller number of relevant options while also allowing them to apply multiple tags if the sentence explicitly includes more than one or there is insufficient evidence to choose only one tag. The guided coding process also allows for the possibility that earlier coarse decisions have less error than later fine-grained decisions.

⁶See the full codebook on Github Repository ICBeEventData.

removing missiles from Turkey. Through compounding, the ontology can capture what players were said to have known, learned, or said about other specific fully described actions.

No existing event data distinguishes thoughts, speeches, and actions. In fact, most only try to code actions and entirely omit thoughts and speech acts despite recognition of their importance in international politics (Smith, 1998). Scholars have opted against coding thoughts and speech acts because of a lack of confidence the full universe could be readily observed and consequently at least theoretically be included.⁷ But the perfect should not be the enemy of the good, and measurement challenges are only overcome after an initial attempt to estimate difficult-to-observe concepts of interest. The ICB narratives are one of the better sources for this endeavor due to the consistent use of high-quality primary source material that takes advantage of qualitative methods well-suited to identifying thoughts and speech acts like archival work and expert interviews.

Each crisis was typically assigned to two expert coders and two novice coders with an additional tie-breaking expert coder assigned to sentences with high disagreement.⁸ For the purposes of measuring intercoder agreement and consensus, we temporarily disaggregate the unit of analysis to the Coder-Crisis-Sentence-Tag ($n = 993,731$), where a tag is any unique piece of information a coder can associate with a sentence such as an actor, date, behavior, etc. We then aggregate those tags into final events ($n = 18,783$), using a consensus procedure (SI Appendix 2.2) that requires a tag to have been chosen by at least one expert coder and either a majority of expert or novice coders. This screens noisy tags that no expert considered possible but leverages novice knowledge to tie-break between equally plausible tags chosen by experts. Requiring sentence-tag matching may underestimate agreement but minimizes the inclusion of noise and allows for additional validation. Once filtered for agreement, we find 472 actors and 119 different behaviors: 12 thought, 13 speech, and 94 actions.

3. Performance comparison

3.1 Internal consistency

We evaluate the internal validity of the coding process in several ways. For every tag applied we calculate the observed intercoder agreement as the percent of other coders who also applied that same tag (SI Appendix 2.3). Across all concepts, the Top 1 Tag Agreement was low among novices (31 percent), moderate for experts (65 percent), and high (73 percent) following the consensus screening procedure.

We attribute the remaining disagreement primarily to three sources. First, we required coders to rate and justify their confidence in the coding. They reported low confidence for 20 percent of sentences; 45 percent of those were due to a mismatch between the ontology and the text (“survey doesn’t fit event”) and 46 percent were from a lack of information or confused writing in the source text (40 percent “more knowledge needed,” 6 percent “confusing sentence”). Observed disagreement varied predictably with self-reported confidence (SI Appendix 2.4). Second, as intended, agreement is higher (75–80 percent) for questions with fewer options near the root of the ontology compared to agreement for questions near the leaves of the ontology (50–60 percent). Third, individual coders exhibit non-trivial coding styles, e.g. some more expressive coders applied many tags per concept while others focused on only the single best match. We further observed unintended synonymity, e.g. the same information can be framed as either a threat to do something or a promise not to do something.

3.2 Improvement over existing efforts

To evaluate our coding process relative to existing datasets, we measure the recall and precision of ICB events in absolute terms and relative to other existing systems. Recall measures the share of desired information recovered by a sequence of coded events while precision measures the degree

⁷Even the coding of overt actions like MIDs is not without contention (Gibler, 2018).

⁸Expert coders were graduate students or postgraduates who collaboratively developed the ontology and documentation for the codebook. Undergraduate coders were students who engaged in classroom workshops.

to which a sequence of events correctly and usefully describes the information in history. To aid in subjective evaluation of the precision and recall of ICBe for each event, we provide full ICB narratives, ICBe coding in an easy-to-read iconographic form, and a wide range of visualizations for every case on the companion website.

Recall for historical episodes is poorly defined for two reasons. History may or may not be written by the victors but by virtue of being written by *someone* there is no genuine ground truth about what occurred, only surviving texts about it (Turberville, 1933). Second, there is no *a priori* guide to what information is necessary detail and what is ignorable trivia. History suffers from what is known as the Coastline Paradox (Mandelbrot, 1983)—it has a fractal dimension greater than one such that the more you zoom in, the more detail you will find about individual events as well as in between any two discrete events. The ICBe ontology is a proposal about what information is important, but we need an independent benchmark to evaluate whether that proposal is a good one and that allows for comparing proposals from event projects that had different goals. We need a yardstick for history.

Our strategy for dealing with both problems is a plausibly objective yardstick called a synthetic historical narrative. We collect a large diverse corpus of narratives spanning timelines, encyclopedia entries, journal articles, news reports, websites, and government documents. Using natural language processing (fully described in SI Appendix 3.1), we identify details that appear across multiple accounts. A detail refers to the smallest textual unit for which we can calculate similarity across corpora to identify whether sentences semantically refer to the same broader observed event (Narayan *et al.*, 2018). The more accounts that mention a detail, the more central it is to understanding the true historical episode. The theoretical motivation is that authors face word limits which force them to pick and choose which details to include, and they choose details that serve the specific context of the document they are producing. With a sufficiently large and diverse corpus of documents, we can vary the context while holding the overall episode constant and see which details tend to be invariant to context. Sufficiently similar details were binned together and then summarized so they could be compared to the coding in ICBe. This presents a harder evaluation baseline than comparing ICBe's recall to just that of ICB since there are non-crisis aspects of these events that may be included in other narratives but are out of the scope of our data. For example, the nationalization of businesses in Cuba may be included as important context in the Cuban Missile Crisis in documents that do not focus on the crisis dimensions like ICB. Using this hard case, a recall measure of ICBe on the synthetic narratives thus serves as a way to evaluate the breadth of ICBe's ontology and potential application to non-crisis international events.

We find substantive variation in recall across existing state of the art methods. Mentions of a detail across accounts are exponentially distributed with context-invariant details appearing dozens to hundreds of times more than context-dependent details.⁹ Furthermore, crisis start and stop dates are arbitrary, and the historical record points to many precursor events as necessary detail for understanding later events. Figure 2 compares ICBe's recall with that of existing datasets for the two case studies detailed in Section 4. ICBe strictly dominates all of the systems but ICEWs in recall though we note that the small sample sizes mean these systems should be considered statistically indistinguishable. Across all existing datasets and ICBe, recall increases with the number of document mentions which is an important sign of validity for both them and our benchmark. The one outlier is Phoenix which in the Cuban Missile Crisis case is so noisy that its recall curve is flat to decreasing as mentions increase. The two episode-level datasets (MIDs and ICM) have

⁹As the ICB narratives are intended to explain conflictual behavior in a political context, many of the missing events concern more economic components of conflict (eg. nationalizing a foreign business). Even when they occur in the context of a crisis, these events largely fall outside the sample of information on which ICBe's ontology is currently trained. Even with this limitation, ICBe is more comprehensive than the existing datasets that do try to code the economic dimensions of these crises. We see expanding the ontology to broader international phenomenon as a promising future implementation of our model.

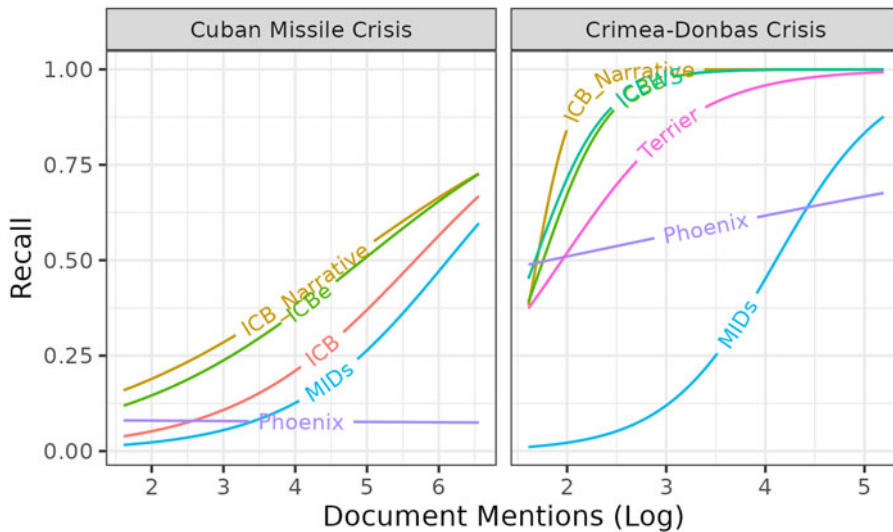


Figure 2. Recall comparison of two cases across existing state of the art efforts. Higher y-axis values represent higher recall and higher x-axis values represent number of times that detail is mentioned across the full corpus used to construct the synthetic narrative.

low coverage of contextual details. The two other dictionary systems ICEWs and Terrier have higher coverage, with ICEWs outperforming Terrier. Importantly our corpus of ICB narratives has high recall of frequently mentioned details giving us confidence in how those summaries were constructed, and ICBe lags only slightly behind showing that it left little additional information on the table.¹⁰

The second component of event measurement validation is precision. It does little good to recall a historical event but too vaguely (e.g., MIDs describes the Cuban Missile Crisis as a blockade, a show of force, and a stalemate) or with too much error to be useful for downstream applications (e.g., ICEWS records 263 “Detonate Nuclear Weapons” events between 1995 and 2019). ICBe’s ontology and coding system are designed to strike a balance so that the most important information is recovered accurately but also abstracted to a level that is still useful and interpretable.

We demonstrate ICBe’s precision in a number of different ways. First, we develop the iconography system for presenting event codings as coherent statements that can be compared side by side to the original source narrative for every case on the companion website. We further provide a stratified sample of event codings alongside their source text (SI Appendix 4.2). We find both the visualizations of macrostructure and head-to-head comparisons of ICBe codings to the raw text to strongly support the quality of ICBe. Second, we develop a visualization we call a crisis map, a directed graph intersected with a timeline. A researcher should be able to lay out the events of a crisis on a timeline and read off the macrostructure of an episode from each individual move. A crisis map using ICBe for the Cuban Missile Crisis case study is provided in Figure 5, crisis maps for the two case studies using existing event datasets can be found in SI Appendix 4.3 and 4.4, and crisis maps for all crises using all datasets can be found on the companion website. The crisis maps reveal episode-level datasets like MIDs or the original ICB are too sparse and vague to reconstruct the structure of the crisis (SI Appendix 4.3 and 4.4). On the other end of the spectrum, the high recall dictionary-based event datasets like Terrier and ICEWs produce

¹⁰Although Figure 2 focuses only on two crises, the synthetic narrative approach and recall comparison can, and should, be more broadly applied to all international crises in a way that could reveal systematic blindspots across datasets.

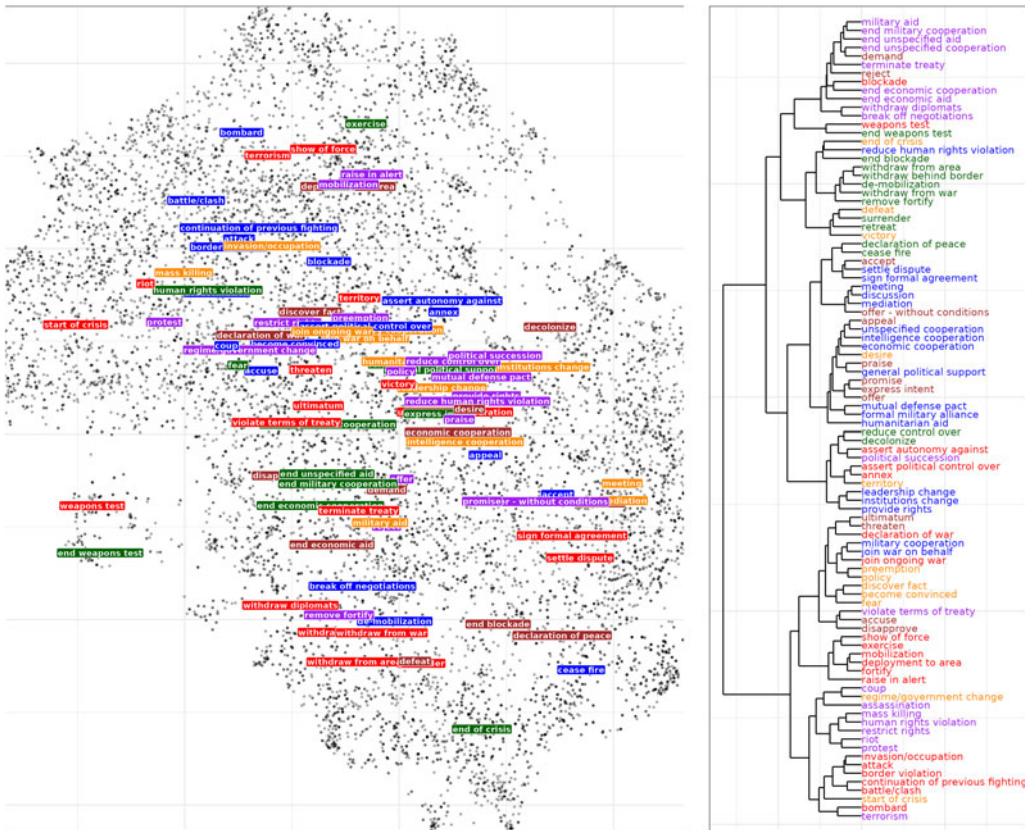


Figure 3. Computational evaluation of the precision of ICBe event codings. The plot on the left is a map of the semantic meaning of every sentence in the corpus (black points) as assigned by a large language model (Paraphrase-MPNET-base-v2) and projected down into two dimensions (UMAP). Overlaid are the median semantic locations of each label assigned by ICBe coders (colored labels). The labels with similar meaning are assigned to sentences with similar semantic meaning, creating an observable structure and pattern we would not observe with low-quality coding where tag location would instead appear random. The plot on the right shows a hierarchical dendrogram clustering labels into groups by their average semantic location with more similar labels being more closely connected on the tree. The clustering by color indicates it closely mirrors the intended ICBe ontology, suggesting high precision in the coding.

so many noisy events (several hundred thousand) that even with heavy filtering their crisis maps are completely unintelligible. Further, because of copyright issues, none of these datasets directly provide the original text spans making event-level precision difficult to verify.

We further want to automatically verify the precision of individual ICBe event codings, which we can do in the case of ICBe because each event is mapped to a specific span of text. Our proposed measure is a reconstruction task to see whether our intended ontology can be recovered through only unsupervised clustering of sentences they were applied to. Figure 3 shows the location of every sentence from the ICBe corpus in semantic space, as embedded using the same large language model as before, and the median location of each ICBe event tag applied to those sentences.¹¹ Labels reflect the individual leaves of the ontology and colors reflect the higher level coarse branch nodes of the ontology. If ICBe has high precision, substantively similar tags ought to have been applied to substantively similar source text, which is what we see both in

¹¹We preprocess sentences to replace named entities with a generic Entity token.

two dimensions in the main plot and via hierarchical clustering on all dimensions in the dendrogram along the right-hand side.¹²

4. Case illustrations

In this section, we focus our validation on two case studies for which we have produced synthetic narratives using the method described in Section 3.2. The first is the Cuban Missile Crisis which took place primarily in the second half of 1962, involved the United States, the Soviet Union, and Cuba, and is widely known for bringing the world to the brink of nuclear war (Figure 1). The second is the Crimea-Donbas Crisis which took place primarily in 2014, involved Russia, Ukraine, and NATO, and within a decade spiraled into a full-scale invasion (SI Appendix 4.1). We choose these cases because they are significant in contemporary international relations, are widely known across academic disciplines as well as among the public, and are sufficiently brief to evaluate in depth. They are similar in that both cases involve a superpower in crisis with a neighbor that changed from a friendly to a hostile regime, both held implications for the economic and military security for the superpower by risking full-scale invasion, and both eventually invited intervention by an opposing superpower.

4.1 Cuban Missile Crisis (1962)

A synthetic historical narrative for the Cuban Missile Crisis appears in Figure 4, with 51 events drawn from 2,020 documents. Each row represents a detail that appeared in at least five documents along with an approximate start date, a handwritten summary, the number of documents it was mentioned in, and whether it could be identified in the text of the original ICB corpus, our ICBe events, and any of the competing existing models.

ICBe's improved recall of the Cuban Missile Crisis relative to the state of the art was summarized in Section 3.2, but the events that explain that improvement can now be seen. Our ground truth ICB narrative contains 17/51 of the events from the synthetic narrative of a case that includes high-level previously classified details. ICBe captures nearly all details included in ICB as well as more details from the synthetic narrative than any competing dataset. Phoenix includes some earlier information than ICBe like the nationalization of businesses and back channel negotiations, but the crisis narrative has a clean canonical end with the Soviets agreeing to withdraw missiles. ICBe stands out in including more communicative behavior (do-speech) than existing datasets like US threats to attack and later promises not to invade. Given the recognized importance of threat credibility for understanding international conflict, the addition of this information is a substantively important improvement over the existing state of the art (Slantchev, 2011).

Figure 5 shows the crisis map for the Cuban Missile Crisis. Looking at the crisis on a timeline, one can now identify the structure of actors and the environment, along with its supporting details, in a way that validates the precision of ICBe. Although harder to measure objectively, this crisis map provides face validity that ICBe's account is not too vague, but also not unnecessarily detailed. We include much of the geopolitically important details like Soviet deployment, US discovery of that deployment, heightened alert levels, a blockade, and negotiations that ended with a formal agreement. At the same time, the crisis map indicates that ICBe does not include unnecessary nuances that preclude useful comparison to other international events.

4.2 Crimea-Donbas (2014)

A synthetic historical narrative for the 2014 Crimea-Donbas Crisis (30 events drawn from 971 documents) appears in Figure 6. As in the earlier case, rows represent details that appeared in at least five documents and whether it is identified in ICBe and existing datasets.

¹²Hierarchical clustering on cosine similarity and with Ward's method.

YMD	Ground Truth Events	Docs	ICB Corpus	ICBe	ICB	MIDs	Phoenix
1958-12-31	communist coup	128					
1959-06-08	nationalizes owned businesses	28					✓
10-26	backchannel negotiates with	18					✓
1960-01-01	begins recon flights over	10					
03-17	prepares for invasion of	159	✓	✓			
05-01	U-2 downed over	8				✓	
07	establishes diplomatic and trade relations with	10					
07-08	embargos	21					
09-14	attempts assassination	9					
1961-01-03	breaks diplomatic ties	40					
04-17	attempts coup in	16					
	invades	192	✓	✓	✓	✓	
19	invasion fails	5					
06-01	provides economic and military aid to	43	✓	✓	✓		
08-13	begins construction of Berlin Wall	62	✓	✓	✓		
11-30	covert destabilization efforts against	44					
1962-04-01	places nuclear missiles in	93					
05-21	begins placing nuclear missiles in	31					
07-01	asks for weapons	5					
	deploy troops to	5					
	places nuclear missiles in as response to placing nuclear missiles in	347					
08-10	begins to suspect nuclear missiles will be placed in	29	✓				
09-04	meets to denies presence of missiles	14	✓				
	demands withdrawal and threatens nuclear response on if attacked from	125					
11	threatens war with if attack on or ships	8					
10-01	deploy nuclear armed submarines	5					
14	discovers nuclear missiles in starts crisis	705	✓	✓	✓		
17	mobilizes troops for invasion of	13	✓	✓			
18	meets to denies presence of missiles	21					
22	blockades	400	✓	✓	✓	✓	
	demands withdrawal	7					
	raises nuclear alert	5					
	threatens military attack	10					
23	OAS statement of support for	7	✓	✓			
	meets denies offensive intention	8					
24	respects blockade of	23			✓		
	raises nuclear alert	14					
25	confronts at the	30					✓
26	offers remove missiles for no invasion pledge	60	✓	✓	✓		
27	nuclear missiles in operational	5					
	accidentally violates airspace	25					
	interdicts submarine	16					
	offers withdrawal of nuclear missiles in for missiles in	115					
	promises to not invade	32	✓	✓			✓
	U-2 shot down over	119	✓	✓			
28	withdraws nuclear missiles in trade for promise to not invade ends crisis	647	✓	✓	✓	✓	
30	refuses observers	7	✓	✓			
11-20	ends blockade	22	✓	✓	✓		
1963-04-01	removes missiles from	5					
08-05	sign Nuclear Test Ban Treaty	94					
30	install hotline	55					

Figure 4. Synthetic narratives combine several thousand accounts of each crisis into a single timeline of events, taking only those mentioned in at least 5 or more documents. Checkmarks represent whether that event could be hand matched to any detail in the ICB corpus, ICBe dataset, or any of the other event datasets (SI Appendix 3.2 and 3.3).

Again quantitatively summarized earlier in Section 3.2 (Figure 2), our ground truth ICB narrative contains 23/30 of the events from the synthetic narrative. Like the gray zone precursor to the Cuban Missile Crisis (Cormac and Aldrich, 2018), Ukraine provided several security guarantees to Russia that were potentially undone, e.g. a long-term lease on naval facilities in Crimea. But unlike the Cuban Missile Crisis, the end of this crisis is unclear, with the event meekly ending with a second cease-fire agreement (Minsk II) but continued fighting. ICBe again recalls more important information about the crisis than any existing dataset, particularly information concerning the behavior of non-state separatist groups like the Donetsk People’s Republic (DPR) and Luhansk People’s Republic (LPR).

As this more recent case reflects primarily public reporting rather than the previously classified details relevant for the Cuban Missile Crisis, ICBe’s improvement relative to the global and real-

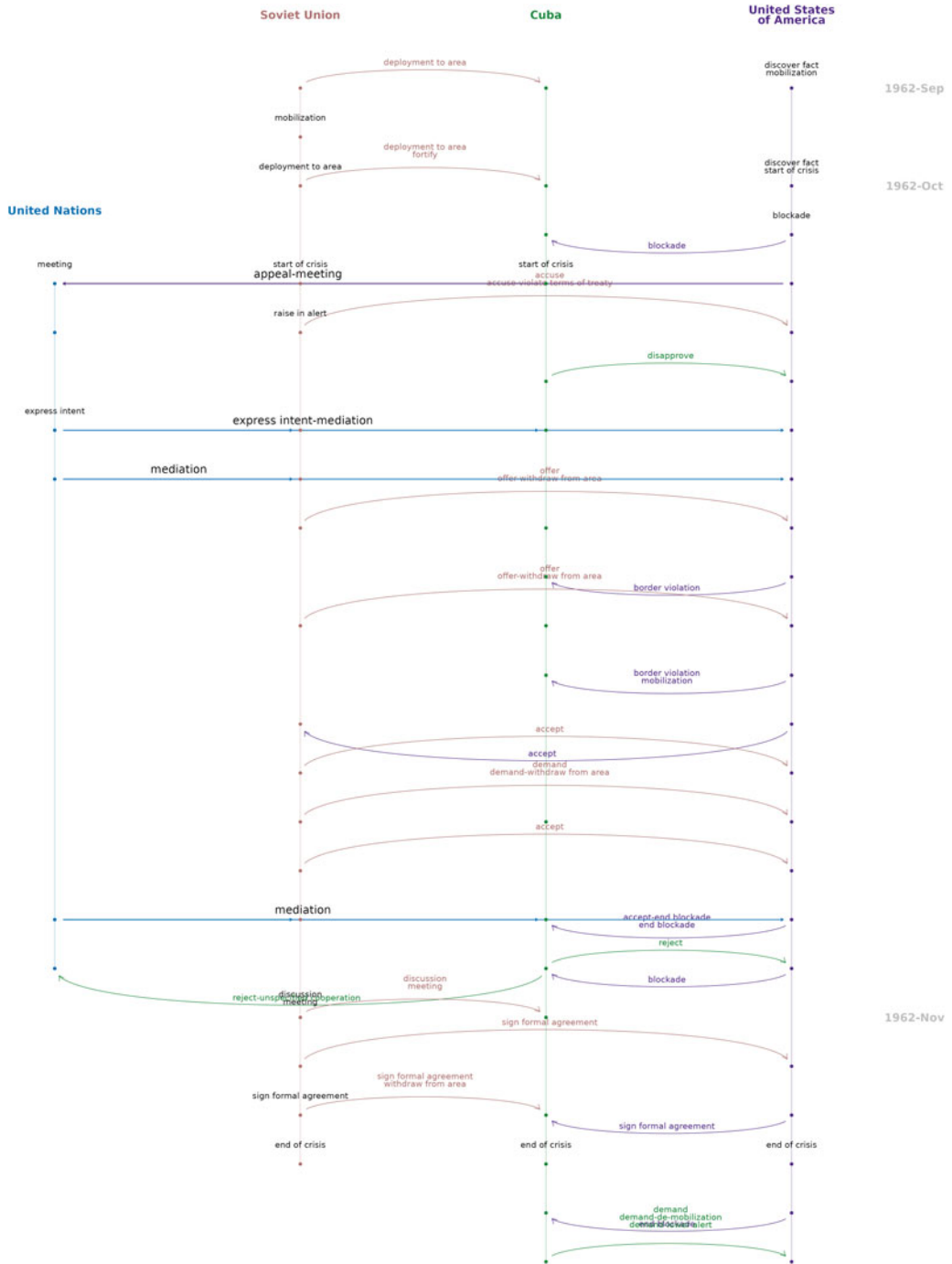


Figure 5. Crisis map for the Cuban Missile Crisis. The start of the crisis is at the top and end of the crisis is at the bottom, with each actor in a column with labeled points identifying their speeches, actions, and thoughts.

time coverage of dictionary-based event systems is still present, but less pronounced. We want to take seriously the possibility that some functional transformation could recover the precision of ICBe. For example, Terechshenko (2020) attempts to correct for the mechanically increasing

YMD	Ground Truth Events	Docs	ICB Corpus	ICBe	MIDs	phoenix	terrier	icews
1994-12-05	Ukraine nuclear disarmed in exchange for Russia promise to never invade	7						
2004-04-02	NATO expands to include Estonia, Latvia, and Lithuania	5						
2008-01-01	NATO provides incredible offer of membership to Ukraine	7	✓	✓				
08-01	Ukraine invades Georgia and annexes Abkhazia and Ossetia and Georgia ratified the Russian Ukrainian Naval Base for Gas	10	✓	✓	✓			
2010-04-21	Treaty, extending the Russian Navy's lease of Crimean facilities for 25 years after 2017	5						
2013-11-01	Ukraine plans to join EU trade agreement	6	✓			✓		✓
21	Ukraine rejects EU trade agreement	42	✓	✓			✓	✓
12-17	Ukraine offers debt relief and discounted energy to EU	6	✓	✓		✓	✓	✓
2014-02-01	EU provide econ and military aid to Ukraine	5	✓	✓		✓	✓	✓
21	Ukraine backs EU political settlement	26	✓	✓		✓	✓	✓
22	Ukraine leader removed and flees to EU	10	✓	✓		✓	✓	✓
27	Ukraine backed gunmen begin seizing government buildings in Crimea	65	✓	✓		✓	✓	✓
03-01	EU votes to deploy military to Ukraine	10	✓	✓		✓	✓	✓
16	Crimea independence referendum	11	✓	✓		✓	✓	✓
16	Russia mobilizes forces to Crimea and EU border	12	✓	✓		✓	✓	✓
17	Russia backed separatists attack EU in Donbas	11	✓	✓		✓	✓	✓
17	EU sanctions Russia	6	✓	✓		✓	✓	✓
18	Russia annexes Crimea	83	✓	✓	✓	✓	✓	✓
04-06	Russia backed separatists begin fighting in EU (Donetsk and Luhansk)	180	✓	✓	✓	✓	✓	✓
23	EU military exercises at EU border	146	✓	✓	✓	✓	✓	✓
05-11	DPR and the LPR declare independence	7	✓	✓		✓	✓	✓
21	DPR requests EU Russian military intervention	5	✓	✓		✓	✓	✓
07-17	Russia shoots down passenger jet (Malaysia Airlines Flight 17)	5	✓	✓		✓	✓	✓
09-05	EU sanctions Russia	6	✓	✓		✓	✓	✓
06	EU econ sanction EU	5	✓	✓		✓	✓	✓
2015-02-12	Ukraine, EU, the DPR and the LPR signed a ceasefire agreement, the Minsk I	10	✓	✓		✓	✓	✓
06	Ukraine, EU, DPR, and the LPR, continue fight with EU	15	✓	✓		✓	✓	✓
21	Ukraine, EU, the DPR and the LPR signed a ceasefire agreement (Minsk II)	10	✓	✓		✓	✓	✓
21	EU recognizes Donetsk People's Republic and the Luhansk People's Republic	6						

Figure 6. Synthetic narratives combine several thousand accounts of each crisis into a single timeline of events, taking only those mentioned in at least five or more documents. Checkmarks represent whether that event could be hand matched to any detail in the ICB corpus, ICBe dataset, or any of the other event datasets (SI Appendix 3.2 and 3.3).

amount of news coverage each year by de-trending violent event counts from Phoenix using a human-coded baseline. Others have focused on verifying precision for ICEWs on specific subsets of details against known ground truths, e.g. geolocation (Cook and Weidmann, 2019), protest events (80 percent) (Wüest and Lorenzini, 2020), and anti-government protest networks (46.1 percent) (Jäger, 2018).

We take the same approach here in Figure 7, selecting four specific CAMEO event codings and checking how often they reflect a true real-world event from the Crimea-Donbas synthetic narrative. We choose four event types around key moments in the crisis. The start of the crisis revolves around Ukraine backing out of a trade deal with the EU in favor of Russia, but “sign formal agreement” events act more like a topic detector with dozens of events generated by discussions of a possible agreement but not the actual agreement which never materialized. The switch is caught by the “reject plan, agreement to settle dispute” event type, but also continues for Viktor Yanukovich even after he was removed from power because of articles retroactively discussing the cause of his removal. Events for “use conventional military force” capture a threshold around the start of hostilities and who the participants were but not any particular battles or campaigns. Likewise, “impose embargo, boycott, or sanctions” captures the start of waves of sanctions and from who but are effectively constant as the news coverage does not distinguish between subtle changes or additions. In sum, dictionary-based methods on news corpora tend to have high recall because they parse everything in the news, but for the same reason, their specificity for most event types is too low to back out individual chess-like sequencing that ICBe aims to record.

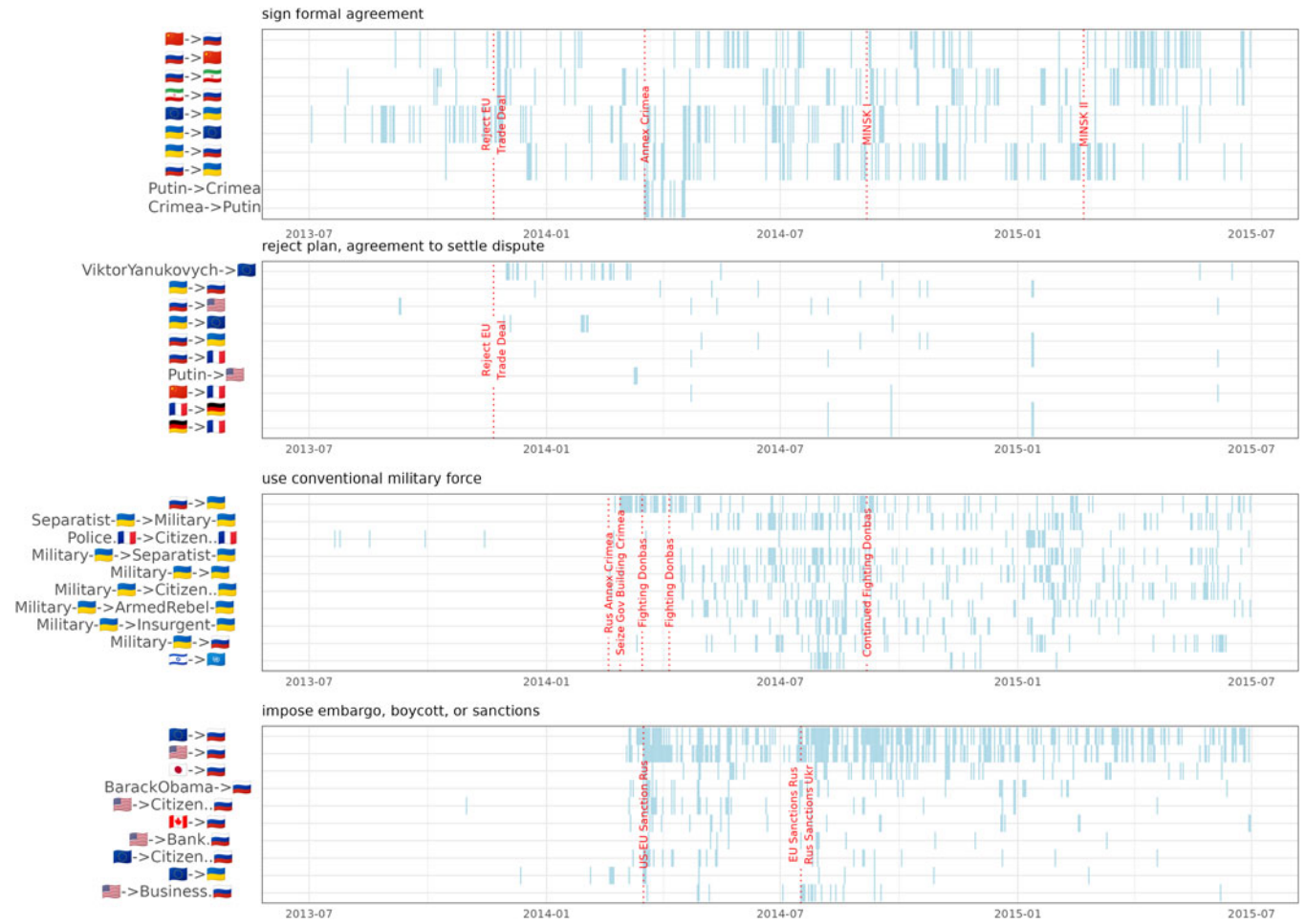


Figure 7. The unit of analysis is the dyad-day. Top 10 most active dyads per category shown. Red text shows events from the synthetic narrative relative to that event category. Blue bars indicate an event recorded by ICEWs for that dyad on that day.

5. Conclusion

The scope and complexity of international politics should not discourage the identification of trends, patterns, and regularities. In undertaking event abstraction from narratives about key historical episodes in international relations, this paper has proposed a mapping between unstructured historical records and a structured ontology of these events with high coverage of concepts of interest. Multiple validity checks find the resulting codings have high internal validity (e.g., intercoder agreement) and external validity (i.e., matching source material in both micro-details at the sentence level and macro-details spanning full historical episodes). Further, these codings perform much better in terms of recall, precision, coverage, and overall coherence in capturing these historical episodes than existing event systems used in international relations.

These data, along with the open-source code, documentation, and companion website provide several substantive and methodological contributions to the discipline. Substantively, these data are appropriate for statistical analysis of hard questions in the study of crises like interactions between means of warfare and the preconditions for conflict escalation (Gannon, 2022). Methodologically, our mapping from codings to source text at the sentence level provide a new resource for natural language processing with access to coder-level disaggregation that furthers the study of uncertainty in the interpretation of international events and in the quantitative coding of historical events. Finally, we provide a companion website (crisisevents.org) that incorporates detailed visualizations of all the data introduced here as a new resource for the study of international crises in a scalable, yet detailed, manner.

Supplementary material. The supplementary material for this article can be found at <https://doi.org/10.1017/psrm.2024.17>. To obtain replication material for this article, <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2F2FDVN%2FMNVUEP&version=DRAFT>

Data. This article's data, supplementary appendix, replication material, and visualizations of every historical episode are available on the GitHub repository ICBEventData and through the companion website crisisevents.org.

Acknowledgements. We thank the ICB Project and its directors and contributors for their foundational work and their help with this effort. We make special acknowledgement of Michael Brecher for helping found the ICB project in 1975, creating a resource that continues to spark new insights to this day. We thank the many undergraduate coders. Thanks to the Center for Peace and Security Studies and its membership for comments. Special thanks to Rebecca Cordell, Philip Schrodt, Zachary Steinert-Threlkeld, and Zhanna Terechshenko for their generous feedback. Thank you to the cPASS research assistants: Helen Chung, Daman Heer, Syeda ShahBano Ijaz, Anthony Limon, Erin Ling, Ari Michelson, Prithviraj Pahwa, Gianna Pedro, Tobias Stodiek, Yiyi 'Effie' Sun, Erin Werner, Lisa Yen, and Ruixuan Zhang.

Author contributions. Conceptualization: R. W. D., E. G., J. L.; methodology: R. W. D., T. L. S.; software: R. W. D.; validation: R. W. D., T. L. S.; formal analysis: R. W. D., T. L. S.; investigation: S. C., R. W. D., J. A. G., C. A., N. L., E. M., J. M. C. N., D. P., D. M. Q., J. W.; data curation: R. W. D., D. M. Q., T. L. S., J. W.; writing — original draft: R. W. D., T. L. S.; writing — review and editing: R. W. D., J. A. G., E. G., T. L. S.; visualization: R. W. D., T. L. S.; supervision: E. G.; project administration: S. C., R. W. D., J. A. G., D. M. Q., T. L. S., J. W.; funding acquisition: E. G., J. L.

Financial support. This work was supported by a grant from the Office of Naval Research N00014-19-1-2491 and from the Charles Koch Foundation 20180481. The financial sponsors played no role in the design, execution, analysis and interpretation of data, or writing of the study.

Competing interest. The authors declare that there are no competing interests.

References

- Allen MA, Flynn ME and Machain CM (2022) US global military deployments, 1950–2020. *Conflict Management and Peace Science* 39(3), 351–370. <https://doi.org/10.1177/07388942211030885>.
- Allison GT and Zelikow P (1971) *Essence of Decision: Explaining the Cuban Missile Crisis*. Vol. 327. Boston: Little, Brown, and Co.
- Althaus S, Bajjalieh J, Carter JF, Peyton B and Shalmon DA (2019) Cline Center Historical Phoenix Event Data Variable Descriptions. *Cline Center Historical Phoenix Event Data*.
- Balali A, Asadpour M and Jafari SH (2021) CoFEE: a Comprehensive Ontology for Event Extraction from Text. arXiv. <https://doi.org/10.48550/arXiv.2107.10326>.

- Beardsley K (2011) *The Mediation Dilemma*. Ithaca and London: Cornell University Press.
- Beardsley K, James P, Wilkenfeld J and Brecher M (2020) The International Crisis Behavior Project. *Oxford Research Encyclopedia of Politics*. <https://oxfordre.com/politics/view/10.1093/acrefore/9780190228637.001.0001/acrefore-9780190228637-e-1638>. <https://doi.org/10.1093/acrefore/9780190228637.013.1638>.
- Beger A, Morgan RK and Ward MD (2021) Reassessing the role of theory and machine learning in forecasting civil conflict. *Journal of Conflict Resolution* **65**, 1405–1426. <https://doi.org/10.1177/0022002720982358>
- Beiler J, Brandt PT, Halterman A, Schrodt PA, Simpson EM and Alvarez R (2016) Generating political event data in near real time: opportunities and challenges. In Alvarez RM (ed.), *Computational Social Science Discovery and Prediction*. Cambridge: Cambridge University Press, pp. 98–120.
- Ben-Yehuda H and MishaliRam M (2006) Ethnic actors and international crises: theory and findings, 1918–2001. *International Interactions* **32**, 49–78.
- Bloomfield LP and Moulton A (1989) CASCON III: Computer-aided System for Analysis of Local Conflicts. MIT Center for International Studies, Cambridge.
- Boschee E, Lautenschlager J, O'Brien S, Shellman S, Starz J and Ward M (2015) ICEWS Coded Event Data. *Harvard Dataverse* 12.
- Brandt PT, D'Orazio V, Holmes J, Khan L and Ng V (2018) Phoenix Real-Time Event Data.
- Brecher M (1999) International studies in the twentieth century and beyond: flawed dichotomies, synthesis, cumulation: ISA presidential address. *International Studies Quarterly* **43**, 213–264.
- Brecher M and Wilkenfeld J (1982) Crises in world politics. *World Politics* **34**, 380–417. <https://doi.org/10.2307/2010324>
- Brecher M and Wilkenfeld J (1997) *A Study of Crisis*. Ann Arbor, MI: University of Michigan Press.
- Brecher M, Wilkenfeld J, Beardsley KC, James P and Quinn D (2021) International Crisis Behavior Data Codebook. Codebook Version 14.
- Brust C-A and Denzler J (2020) Integrating domain knowledge: using hierarchies to improve deep classifiers. *arXiv:1811.07125 [Cs]*, January. <https://arxiv.org/abs/1811.07125>.
- Bush SS and Hadden J (2019) Density and decline in the founding of international NGOs in the United States. *International Studies Quarterly* **63**, 1133–46. <https://doi.org/10.1093/isq/sqz061>
- Carafano JJ (2014) Measuring military power. *Strategic Studies Quarterly* **8**, 11–18. <https://www.jstor.org/stable/26270616>
- Carter DB (2010) The strategy of territorial conflict. *American Journal of Political Science* **54**, 969–987. <https://doi.org/10.1111/j.1540-5907.2010.00471.x>
- Chenoweth E, Hendrix CS and Hunter K (2019) Introducing the nonviolent action in violent contexts (NVAVC) dataset. *Journal of Peace Research* **56**, 295–305. <https://doi.org/10.1177/0022343318804855>
- Cook SJ and Weidmann NB (2019) Lost in aggregation: improving event analysis with report-level data. *American Journal of Political Science* **63**, 250–264.
- Cormac R and Aldrich RJ (2018) Grey is the new black: covert action and implausible deniability. *International Affairs* **94**, 477–494. <https://doi.org/10.1093/ia/iyy067>
- Davies S, Pettersson T and Öberg M (2022) Organized violence 1989–2021 and drone warfare. *Journal of Peace Research* **59**, 593–610.
- Eck K and Hultman L (2007) One-sided violence against civilians in war: insights from new fatality data. *Journal of Peace Research* **44**, 233–246. <https://doi.org/10.1177/0022343307075124>
- Fazal TM (2011) *State Death: The Politics and Geography of Conquest, Occupation, and Annexation*. Princeton, NJ: Princeton University Press.
- Felbermayr G, Kirilakha A, Syropoulos C, Yalcin E and Yotov YV (2020) The global sanctions data base. *European Economic Review* **129**, 103561. <https://doi.org/10.1016/j.eurocorev.2020.103561>
- Fortna VP (2018) *Peace Time*. Princeton, NJ: Princeton University Press.
- Frederick BA, Hensel PR and Macaulay C (2017) The issue correlates of war territorial claims data, 1816–2001. *Journal of Peace Research* **54**, 99–108. <https://doi.org/10.1177/0022343316676311>
- Gannon JA (2022) One if by land, and two if by sea: cross-domain contests and the escalation of international crises. *International Studies Quarterly* **66**(4), sqac065. <https://doi.org/10.1093/isq/sqac065>
- Gannon JA, Gartzke E, Lindsay JR and Schram P (2024) The shadow of deterrence: why capable actors engage in contests short of war. *Journal of Conflict Resolution* **68**(2-3), 230–268.
- Gartzke E and Lindsay JR (2019) *Cross-Domain Deterrence: Strategy in an Era of Complexity*. Oxford: Oxford University Press.
- Gavin FJ (2014) History, security studies, and the july crisis. *Journal of Strategic Studies* **37**, 319–331. <https://doi.org/10.1080/01402390.2014.912916>
- Gibler DM (2018) *International Conflicts, 1816–2010: Militarized Interstate Dispute Narratives*. Lanham, MD: Rowman & Littlefield.
- Gibler DM and Sarkees MR (2004) Measuring alliances: the correlates of war formal interstate alliance dataset, 1816–2000. *Journal of Peace Research* **41**, 211–222. <https://doi.org/10.1177/0022343304041061>
- Glaser CL (2000) The causes and consequences of arms races. *Annual Review of Political Science* **3**, 251–276. <https://doi.org/10.1146/annurev.polisci.3.1.251>

- Goemans HE, Gleditsch KS and Chiozza G (2009) Introducing archigos: a dataset of political leaders. *Journal of Peace Research* 46, 269–283.
- Goertz G and Diehl PF (1986) Measuring military allocations: a comparison of different approaches. *Journal of Conflict Resolution* 30, 553–581. <https://doi.org/10.1177/0022002786030003009>
- Goldgeier J and Tetlock P (2001) Psychology and international relations theory. *Annual Review of Political Science* 4, 67–92. <https://doi.org/10.1146/annurev.polisci.4.1.67>
- Grant C, Halterman A, Irvine J, Liang Y and Jabr K (2017) OU Event Data Project, December.
- Haffar W (2002) Emergent peacemakers: cataloguing new patterns of activity in post-cold war conflict. *Peace Economics, Peace Science and Public Policy* 8(2), 1–42. <https://doi.org/10.2202/1554-8597.1054>
- Hermann C (1984) Comparative Research on the Events of Nations (CREON) Project: foreign policy events, 1959–1968: Version 1. ICPSR - Interuniversity Consortium for Political and Social Research. <https://doi.org/10.3886/ICPSR05205.V1>.
- Hewitt J (2001) Engaging international data in the classroom: using the ICB interactive data library to teach conflict and crisis analysis. *International Studies Perspectives* 2, 371–383. <https://doi.org/10.1111/1528-3577.00066>
- Holsti OR (1965) The 1914 case. *The American Political Science Review* 59, 365–378. <https://doi.org/10.2307/1953055>
- Hsu A, Höhne N, Kuramochi T, Vilarino V and Sovacool BK (2020) Beyond states: harnessing sub-national actors for the deep decarbonisation of cities, regions, and businesses. *Energy Research & Social Science* 70, 101–738. <https://doi.org/10.1016/j.erss.2020.101738>
- Iakhnis E and James P (2019) Near crises in world politics: a new dataset. *Conflict Management and Peace Science* 38(2), 224–243. <https://doi.org/10.1177/0738894219855610>
- Jäger K (2018) The limits of studying networks with event data: evidence from the ICEWS dataset. *Journal of Global Security Studies* 3, 498–511.
- Jervis R (1978) Cooperation under the security dilemma. *World Politics* 30, 167–214. <https://doi.org/10.2307/2009958>
- Kang DC and Lin AY-T (2019) US bias in the study of Asian security: using Europe to study Asia. *Journal of Global Security Studies* 4, 393–401.
- Kinne BJ (2020) The defense cooperation agreement dataset (DCAD). *Journal of Conflict Resolution* 64, 729–755. <https://doi.org/10.1177/0022002719857796>
- Lacina B (2006) Explaining the severity of civil wars. *Journal of Conflict Resolution* 50, 276–289. <https://doi.org/10.1177/0022002705284828>
- LaFree G and Dugan L (2007) Introducing the global terrorism database. *Terrorism and Political Violence* 19, 181–204.
- Lai B (2004) The effects of different types of military mobilization on the outcome of international crises. *Journal of Conflict Resolution* 48, 211–229.
- Leeds BA (1999) 2003. Alliance reliability in times of war: explaining state decisions to violate treaties. *International Organization* 57, 801–827. <https://doi.org/10.1017/S0020818303574057>
- Leeds BA (1999) Domestic political institutions, credible commitments, and international cooperation. *American Journal of Political Science* 43, 979–1002. <https://doi.org/10.2307/2991814>
- Leng RJ and Singer J (1988) Militarized interstate crises: the BCOW typology and its applications. *International Studies Quarterly* 32, 155–173. <https://doi.org/10.2307/2600625>
- Li Q, Peng H, Li J, Hei Y, Sun R, Sheng J and Guo S (2021) A comprehensive survey on schema-based event extraction with deep learning. [arXiv:2107.02126 \[Cs\]](https://arxiv.org/abs/2107.02126), August. <https://arxiv.org/abs/2107.02126>.
- Lindsay JR and Gartzke E (2020) Politics by many other means: the comparative strategic advantages of operational domains. *Journal of Strategic Studies* 0, 1–34. <https://doi.org/10.1080/01402390.2020.1768372>
- Lupton DL (2018) Reexamining reputation for resolve: leaders, states, and the onset of international crises. *Journal of Global Security Studies* 3, 198–216. <https://doi.org/10.1093/jogss/ogy004>
- Mandelbrot BB (1983) *The fractal geometry of nature*. New York: Freeman.
- McClelland C (1978) World Event/Interaction Survey, 1966–1978. *WEIS Codebook ICPSR* 5211.
- McNabb Cochran K and Long SB (2017) Measuring military effectiveness: calculating casualty loss-exchange ratios for multilateral wars, 1816–1990. *International Interactions* 43, 1019–1040. <https://doi.org/10.1080/03050629.2017.1273914>
- Merritt RL (1994) Measuring events for international political analysis. *International Interactions* 20, 3–33.
- Miller GA (1995) WordNet: a lexical database for English. *Communications of the ACM* 38, 39–41. <https://doi.org/10.1145/219717.219748>
- Min E (2021) Interstate war battle dataset (1823–2003). *Journal of Peace Research* 58, 294–303. <https://doi.org/10.1177/0022343320913305>
- Moyer JD, Turner SD and Meisel CJ (2021) What are the drivers of diplomacy? Introducing and testing new annual dyadic data measuring diplomatic exchange. *Journal of Peace Research* 58(6), 1300–1310. <https://doi.org/10.1177/0022343320929740>
- Narang V and Talmadge C (2018) Civil-military pathologies and defeat in war: tests using new data. *Journal of Conflict Resolution* 62(7), 1379–1405. <https://doi.org/10.1177/0022002716684627>

- Narayan S, Cohen SB and Lapata M (2018) Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. [arXiv:1808.08745](https://arxiv.org/abs/1808.08745).
- O'Neill B (2018) International negotiation: some conceptual developments. *Annual Review of Political Science* 21, 515–533. <https://doi.org/10.1146/annurev-polisci-031416-092909>
- Owsiak AP, Cuttner AK and Buck B (2018) The international border agreements dataset. *Conflict Management and Peace Science* 35, 559–576. <https://doi.org/10.1177/0738894216646978>
- Paige GD (1968) *The Korean Decision, June 24–30, 1950*. New York, NY: Free Press.
- Palmer G, McManus RW, D'Orazio V, Kenwick MR, Karstens M, Bloch C, Dietrich N, Kahn K, Ritter K and Soules MJ (2022) The MID5 dataset, 2011–2014: procedures, coding rules, and description. *Conflict Management and Peace Science* 39(4), 470–482. <https://doi.org/10.1177/0738894221995743>
- Powell R (2002) Bargaining theory and international conflict. *Annual Review of Political Science* 5, 1–30. <https://doi.org/10.1146/annurev.polisci.5.092601.141138>
- Powell JM and Thyne CL (2011) Global instances of coups from 1950 to 2010: a new dataset. *Journal of Peace Research* 48, 249–259. <https://doi.org/10.1177/0022343310397436>
- Quinn D, Wilkenfeld J, Smarick K and Asal V (2006) Power play: mediation in symmetric and asymmetric international crises. *International Interactions* 32, 441–470. <https://doi.org/10.1080/03050620601011107>
- Raleigh C, Linke A, Hegre H and Karlsen J (2010) Introducing ACLED: an armed conflict location and event dataset: special data feature. *Journal of Peace Research* 47, 651–660.
- Ramsay KW (2017) Information, uncertainty, and war. *Annual Review of Political Science* 20, 505–527. <https://doi.org/10.1146/annurev-polisci-051215-022729>
- Reiter D (2015) Should we leave behind the subfield of international relations?. *Annual Review of Political Science* 18, 481–499. <https://doi.org/10.1146/annurev-polisci-053013-041156>
- Reiter D, Stam AC and Horowitz MC (2016) A revised look at interstate wars, 1816–2007. *Journal of Conflict Resolution* 60, 956–76. <https://doi.org/10.1177/0022002714553107>
- Sarkees MReid and Wayman F (2010) *Resort to War: 1816–2007*. Washington, DC: CQ Press.
- Schrodt PA and Hall B (2006) Twenty years of the Kansas event data system project. *The Political Methodologist* 14, 2–8.
- Sechser TS (2011) Militarized compellent threats, 1918–2001. *Conflict Management and Peace Science* 28, 377–401. <https://doi.org/10.1177/0738894211413066>
- Sherman FL (2000) SHERFACS: a cross-paradigm, hierarchical, and contextually-sensitive international conflict dataset, 1937–1985: Version 1. ICPSR – Interuniversity Consortium for Political and Social Research. <https://doi.org/10.3886/ICPSR02292.V1>.
- Slantchev BL (2011) *Military Threats: The Costs of Coercion and the Price of Peace*. Cambridge: Cambridge University Press.
- Smith A (1998) International crises and domestic politics. *American Political Science Review* 92, 623–638. <https://doi.org/10.2307/2585485>
- Spruyt H (1996) *The Sovereign State and Its Competitors: An Analysis of Systems Change*. Princeton, NJ: Princeton University Press.
- Stein AA and Russett BM (1980) Evaluating war: outcomes and consequences. In *Handbook of Political Conflict: Theory and Research*, 399–422. Free Press New York.
- Steinert-Threlkeld ZC (2019) The future of event data is images. *Sociological Methodology* 49, 68–75. <https://doi.org/10.1177/0081175019860238>
- Sullivan PL (2007) War aims and war outcomes: why powerful states lose limited wars. *Journal of Conflict Resolution* 51, 496–524. <https://doi.org/10.1177/0022002707300187>
- Sundberg R and Croicu M (2016) UCDP GED Codebook Version 5.0. Department of Peace and Conflict Research, Uppsala University.
- Sundberg R and Melander E (2013) Introducing the UCDP georeferenced event dataset. *Journal of Peace Research* 50, 523–532.
- Terechshenko Z (2020) Hot under the collar: a latent measure of interstate hostility. *Journal of Peace Research* 57, 764–776. <https://doi.org/10.1177/00223433200962546>
- Trager RF (2016) The diplomacy of war and peace. *Annual Review of Political Science* 19, 205–228. <https://doi.org/10.1146/annurev-polisci-051214-100534>
- Turberville A (1933) History objective and subjective. *History* 17, 289–302. <https://www.jstor.org/stable/24400365>
- Ward MD, Metternich NW, Dorff CL, Gallop M, Hollenbach FM, Schultz A and Weschle S (2013) Learning from the past and stepping into the future: toward a new generation of conflict prediction. *International Studies Review* 15, 473–490.
- Wilkenfeld J and Brecher M (2000) Interstate crises and violence: twentieth-century findings. In Midlarsky MI (ed.), *Handbook of War Studies II*. Ann Arbor, MI: University of Michigan Press, pp. 271–300.
- Wüest B and Lorenzini J (2020) External Validation of Protest Event Analysis. In Kriesi H, Lorenzini J, Wüest B and Hausermann S (eds), *Contention in Times of Crisis: Recession and Political Protest in Thirty European Countries*. Cambridge: Cambridge University Press, pp. 49–74.

- Yarhi-Milo K** (2013) In the eye of the beholder: how leaders and intelligence communities assess the intentions of adversaries. *International Security* **38**, 7–51. https://doi.org/10.1162/ISEC/_a/_00128
- Yarhi-Milo K, Lanoszka A and Cooper Z** (2016) To arm or to ally? The patron's dilemma and the strategic logic of arms transfers and alliances. *International Security* **41**, 90–139. https://doi.org/10.1162/ISEC/_a/_00250
- Zartman I and Faure GO** (2005) *Escalation and Negotiation in International Conflicts*. Cambridge: Cambridge University Press.
- Zhang H and Pan J** (2019) CASM: a deep-learning approach for identifying collective action events with text and image data from social media. *Sociological Methodology* **49**, 1–57. <https://doi.org/10.1177/0081175019860244>