This codebook summarizes the methods used to create the Protest in the Arab Awakening dataset. The focus is all public protests and police actions that occurred in the 5 years prior to and during the Arab Spring in 2010-11 in 22 Middle Eastern states and territories (see list below) between January 1, 2006 and December 31, 2011. Since our cutoff is 2011, we do not deal separately with the Sudan and South Sudan partition but treat these as the single unified “Sudan.” Our event framework includes multiple types of public protest, i.e. demonstrations, strikes, riots and violent protests that occurred in public settings along with police and internal security coercive actions targeted at protestors, insurgents and other dissidents. These are coded from news stories in Reuters World Newswires, which was accessed through Factiva online at the Ohio State University Library System between April 1, 2012 and April 15, 2013. Searches were conducted through the use of a set of keywords designated for each of the five protest event forms (see below) in combination with the names of the 22 states and territories and the year. These searches identified relevant news stories which were then reviewed and used to code specific events, including a set of governmental and policing events in terms of date, actor, event form, target, size estimate and, for protest events, whether there was a coercive police, internal security or military response. This generated a daily event file of 6,862 events. We use a conventional 24-hour clock to identify the date of these events, treating events that occur on multiple days in succession as a single event. In terms of location, we record only the country or territory name as reported in the news story. Because the duration of events cannot consistently be coded from Reuters news stories, we do not include event duration as a variable. For events reported as occurring in multiple locations (e.g. cities, villages, regions of countries) simultaneously during the same 24-hour day, we coded these as single events so long as the actor, event form and target are identical. Because we cannot consistently distinguish between events in Gaza vs. West Bank locations, we treat these together as “Palestine” or “Palestinian territories.” For trans-border events that cross territorial borders (e.g. “Palestinian insurgents fired rockets into Israel today”), the location of the target determines the event location.

Acknowledgement:

This dataset was created with the support of the National Science Foundation under project “NSCC/SA: Deciphering Civil Conflict in the Middle East” (NSCC/SA 0904731).

Citation:

In citing this dataset, please use the following:

**Countries and Territories included:** Algeria, Bahrain, Cyprus, Egypt, Iran, Iraq, Jordan, Kuwait, Lebanon, Libya, Mauritania, Morocco, Oman, Palestine (Gaza & West Bank treated as one unit), Qatar, Saudi Arabia, Sudan, Syria, Tunisia, Turkey, United Arab Emirates (UAE) and Yemen.

**Event Forms:**

Note: In the searches for relevant news stories, we used keywords exclusively for the civil protest and violence events listed to identify relevant news stories. Hence all governmental or policing events are provided from this set of “civil contention” keywords used to locate relevant news stories. No distinct keywords for these governmental events were searched. As a result, it is best to consider all governmental events as reactive to civil contention events.

- **ARES** = Arrests, police raids, police detentions, police seizures of property and persons
- **ATTACK** = Armed attacks using coercive force by police or military (i.e. bombings, air raids, shooting into crowds). This excludes armed attacks by non-state-sanctioned insurgent groups
- **BAN** = Bans and other governmental restrictions on public gathering and expressions of opinion, curfews and martial law
- **CLASH** = Armed clashes between groups where initiation cannot be determined from the news story; the first group mentioned in the story is treated as the “actor” and the second as the “target” but these can be reversed given the reciprocity involved in the reported events; these may involve rival insurgent or dissident groups
- **DEMO** = Protest marches and demonstrations
- **RIOT** = Riots and other crowd like gatherings involving violence against property and/or people. These are events explicitly referenced as “riots”
- **STRIKE** = Strikes (both political and economic)
- **SURV** = Police surveillance, monitoring of dissidents
- **VIOL** = Violent protests. These events are not explicitly referenced as riots and include insurgent attacks, bombings, specific attacks on property and individuals.

**Actor/Target:**

We use the same list of social agents to code both actors and targets. In practice, actors were often coded with greater precision as to the specific agents involved while targets often had to be coded with more inclusive generic labels (such as “government agents”). The following labels borrow on the IDEA framework agents or “source” coding categories (www.vra.net/idea)
ARTS = Artists (musicians, actors, authors, poets and stage performers, but not their unions)
CHRT = Agents explicitly identified as Christian individuals or groups
CIVI = Civic and social groups
CRIM = Agents explicitly identified as criminals
DIPL = Diplomatic officials, (ambassadors and envoys)
EDUC = Public and private educators and administrators of educational institutions
ETHN = Individuals and groups identified by ethnicity
GAGE = Government agents (all agencies, branches, officials, including military—treat as a default category)
INSU = Agents explicitly identified as insurgents (i.e. armed rebels, revolutionaries, terrorists, guerrilla groups, paramilitaries)
JUDI = Judges and court officials
KURD = Agents explicitly identified as Kurdish
MASS = Mass media actors (including journalists, producers and publications)
MED = Health care professionals and institutions
MIGR = Agents explicitly identified as immigrants, refugees and displaced/homeless people
MILI = Official armed forces, military police and border guards
MOSL = References to Moslem individuals or groups
NLEG = Representatives, councils and parliaments
OCCU = Agents referred to by their occupation (unless identified by a more specific label)
OPPS = Agents referred to as political opposition groups and individuals
POLI = Law enforcement agencies, officials, prosecutors, prison guards, police involved in crowd control but not military police or border guards
RELI = Religious leaders, groups, organizations and individuals (unless identified by a more specific label)
SHIITE = Individuals and groups identified as Shiites (including insurgents & dissidents)
STUD = Students and student groups
SUNNI = Individuals identified as Sunnis (including insurgents & dissidents)
UNIO = Unions and worker advocacy groups
UNKC = Unknown agents that represent civil society actors in some capacity
UNKS = Unknown agents that represent state actors in some capacity
US = United States military
YOUTH = Young adults and children not identified as students

**Protest Size:**

The size of events is coded only for civilian contentious events (“DEMO”, “STRIKE,” “RIOT” and “VIOL”). We use a “.” to indicate events where there was insufficient evidence to determine size. The following code is used to identify these event participation levels:

1 = less than 10
2 = less than 100
3 = less than 1000
4 = less than 10,000
5 = less than 100,000
6 = less than 1,000,000
7 = over 1,000,000

**Repression:**

Each event conducted is coded yes/no (i.e. 1 or 0) as to whether there was a police or internal security response that involved coercion or violence against protestors, strikers, insurgents and other civilian actors. This is coded as an event attribute. In addition, this responding policing act is also entered into the dataset as a separate event line with the police, military or other governmental actor and the protesters/dissidents treated as targets. This allows “same day” repression events to be linked together. Presence of repression is coded as:

1 = yes
0 = no

**Keyword Search Methodology:**

We use the FACTIVA database (http://global.factiva.com) to perform keyword searches across all Reuters World Newswire articles from 2006 to 2011.

**Search Criteria**

All search terms are listed below. Each country had specific exclusion criteria as listed below in parentheses. Words listed in quotation marks are specific word exclusions used in the searches; while industry exclusions used the built-in search function "Industry." We used the built-in FACTIVA exclusion terms: “exclude by Industry” to eliminate irrelevant business news that used contentious events language. We used the date range function to restrict the date search to, e.g. “01/01/2006 to 12/31/2006.” For each country or territory, we used the built in search function "Region" to identify the country or territory name.

**Source:**

We used the "Source" function to restrict all searches to only "Reuters News" articles. By doing this we exclude the Dow Jones Newswires, Major News and Business Publications, Press Release Wires, and The Wall Street Journal, all of which FACTIVA searches by default. We further restricted language results to "English."

**Region List:**

Algeria
Bahrain
Cyprus
Egypt
Iran (Exclude "Atomic", Exclude by Industry - Energy, Banking/Credit)
Iraq
Jordan
Kuwait (Exclude "Korea"; Exclude by Industry - Banking/Credit, Energy)
Lebanon
Libya (Exclude by Industry - Crude Oil/Natural Gas)
Mauritania
Morocco
Oman
Qatar (Exclude by Industry - Chevron, Crude Oil)
Saudia Arabia (Exclude by Industry - Energy, Business/Consumer Service, Banking/Credit, Food Products, Construction/Real Estate, Telecommunications)
Sudan (Exclude by Industry - Energy; Exclude by Region - Chad)
Syria (Exclude by Industry - Energy, Transportation/Shipping)
Tunisia
Turkey (Exclude by Industry - Energy, Business/Consumer Services, Food Products, Leisure/Arts, Automobiles, Banking/Credit, Transport/Shipping, metals/mining)
UAE (Exclude by Industry - Transportation/Shipping, Energy, Business/Consumer Services, Banking/Credit, Construction/Real Estate)
Yemen (Exclude by Industry - Crude Oil/Natural Gas, Petrol/Coal; Exclude by Region - Saudia Arabia, Syria)

Event Form Search Terms:

As noted earlier, all keywords used in our searches focused on terms relevant to the major forms of civil contention (CLASH, DEMO, RIOT, STRIKE, VIOL). This means no keywords were entered for the 4 governmental repression events (ARES, ATTACK, BAN, SURV). We conducted four searches per country/territory per a year corresponding to each type of civilian contention event form. The following search terms were entered into the FACTIVA search form "At least one of these words" and exclusion criteria (listed below) were entered into the FACTIVA search form "None of these words." The default word search and exclusionary terms for each event form are as follows:

CLASH (Armed clashes between dissidents and government actors)
"clash" OR "clashes" OR "clashing"

DEMO (Demonstration):
"protest" OR "demonstrate" OR "petition" OR "sit-in" OR "march" OR "picket" OR "parade" OR "occupying" OR "leaflet" OR "hunger strike" OR "set themselves on fire" OR "blockade" OR "blockading" OR "procession"
RIOT (Riot):
"riot" OR "mobs" OR "melee" OR "looting" OR "looters" OR "ransacking"

STRIKE (Strike):
"strike" OR "striking" OR "striker" OR "boycott"

VIOL (Violent Protest):
"violence" OR "bomb" OR "bombing" OR "attack" OR "skirmish" OR "battles" OR
"battling" OR "assassin" OR "assassinate" OR "seizures" OR "hijacking" OR "coup" OR
"mutiny" OR "explosion" OR "uprising" OR "rebel"

Exclusion Terms:

In addition a number of terms were used to exclude non-news story Reuters texts that are not useful for coding events. The following exclusion terms were used to limit our reviews of Reuters texts:

"analysis" OR "interview" OR "factbox" OR "Q&A" OR "table" OR "press digest" OR
"FEATURE-" OR "preview-" OR "chronology-" OR "stocks" OR "Reuters Middle East
Highlights"

Missing Data:

We use a “.“ (missing) to indicate country/year combinations where no events were found or when an event was found but there was insufficient information to properly code the actor, target, event type and the size of the event.