Of flour, yeast and water

A recipe for (Syrian) conflict data

Clingendael Report

Kars de Bruijne

Clingendael
Netherlands Institute of International Relations

Clingendael

Netherlands Institute of International Relations

Of flour, yeast and water

A recipe for (Syrian) conflict data

Kars de Bruijne

Clingendael Report
January 2018

Disclaimer: This report was commissioned by the Netherlands' ministries of Foreign Affairs and Defence within the PROGRESS framework agreement, lot 5, 2017. Responsibility for the contents and for the opinions expressed rests solely with the authors. Publication does not constitute an endorsement by the Netherlands' ministries of Foreign Affairs and Defence.

This report is part of Clingendael's political violence program on information gathering, analysis and recognition of (data) bias by policy-makers.

January 2018

About the authors

Kars de Bruijne is a Research Fellow at Clingendael and a Senior Researcher in the Armed Conflict Location & Event Data Project (ACLED). His research focuses on intrastate conflict, fragile states, trend analyses and expert forecasting.

The Clingendael Institute
P.O. Box 93080
2509 AB The Hague
The Netherlands

Follow us on social media
@clingendaelorg
The Clingendael Institute
The Clingendael Institute

Email: info@clingendael.org
Website: www.clingendael.org

Contents

Introduction

Conflict data is prone to substantial bias. Conflict researchers, for example, have little insight into the exact amount of a conflict that is captured in many databases.[1] Moreover, as conflict data rests on qualitative reporting, it reproduces a number of the biases involved in qualitative reporting. Journalists tend to stay within cities and report less on areas that are harder to access,[2] mobile phone coverage improves reporting[3] but is not available everywhere and the media and local organisations' networks are influenced by their allegiance to parties in conflict.[4] All such problems end up in conflict databases upon which – subsequently – trends are identified which may or may not be real.

This paper provides practical guidance and a *proof of concept* on how to create reliable data. It takes into account the conflict in Syria, one of the most complicated conflicts to date in terms of data gathering. Data was obtained from thirteen public and private organiations, both local and national. The paper provides a recipe on how to combine these data into a reliable conflict database: what ingredients and how many of them should be used to produce reliable data? How much flour, yeast and water – as a figure of speech – do we need in the Syria case? How can this proof of concept be used elsewhere?

Data is based on eleven months of data collection in and on Syria using information generated by thirteen organisations selected out of a much larger number of thirty-five data-providers on Syria. To enable the creation of the data, many directors of the initiatives were directly involved. Eight research analysts sifted through the piles of evidence to create data. Subsequently, three (senior) researchers analysed it. To our

1 Kars de Bruijne and Erwin Van Veen, "Pride and Prejudice: Recognising Bias for Policy-Makers" (The Hague: Clingendael Institute, 2017); Nils B. Weidmann, "A Closer Look at Reporting Bias in Conflict Event Data," *American Journal of Political Science* 60, no. 1 (January 1, 2016): 206–18, https://doi.org/10.1111/ajps.12196.

2 Stathis N. Kalyvas, "The Urban Bias in Research on Civil Wars," *Security Studies* 13, no. 3 (March 2004): 160–90, https://doi.org/10.1080/09636410490914022.

3 Allan Dafoe and Jason Lyall, "From Cell Phones to Conflict? Reflections on the Emerging ICT–political Conflict Research Agenda," *Journal of Peace Research* 52, no. 3 (May 1, 2015): 401–13, https://doi.org/10.1177/0022343314563653; Mihai Croicu and Joakim Kreutz, "Communication Technology and Reports on Political Violence Cross-National Evidence Using African Events Data," *Political Research Quarterly*, September 29, 2016, https://doi.org/10.1177/1065912916670272.

4 Christian Davenport and Patrick Ball, "Views to a Kill: Exploring the Implications of Source Selection in the Case of Guatemalan State Terror, 1977-1995," *Journal of Conflict Resolution* 46, no. 3 (2002): 427–450.

knowledge this project provides the most comprehensive initiative to date to create reliable conflict data.[5] The proof of concept will be used by the Armed Conflict Location & Event Data Project (ACLED) and is funded by the US State Department. Data is freely available.[6]

Research questions:

1. How do different sources of data on the Syrian conflict report on a) the geographical distribution of violence; b) the types of violence; c) the coverage of distinct actors?
2. How can these sources be combined and built upon to create a reliable and consistent conflict database on the Syrian conflict?

This report proceeds as follows: the first section details methodological choices such as the selection of the thirteen local partners, the Syrian context and the research design and process (the section can be skipped for those less interested in methodology). The second section presents the main results. It explains which organisations cover what location, type of actor and type of violence and highlights the consequences for policy-makers. The overall conclusion is as expected: there are major differences in reporting between the initiatives and a reliable database requires a well-designed strategy. The third section explains how the databases can be combined into one reliable set of data and creates it.

5 For related attempts restricted to human rights violations see the work of HRDAG (www.hrdag.org) & Yoshiko M. Herrera and Devesh Kapur, "Improving Data Quality: Actors, Incentives, and Capabilities," *Political Analysis* 15, no. 4 (2007): 365–86.

6 And can be used in policy initiatives.

1 Motivation, partners and methods

The Syrian conflict is perhaps one of the best monitored conflicts of all times. A review of a large number of organisations reporting on the conflict flagged over 30 separate initiatives each collecting information on violence in Syria (see Annex 1). These initiatives can roughly be divided into those which collect or report on human rights violations (Amnesty International, Human Rights Watch, etc.) and those which collect general information on the conflict. In Syria, human rights violation monitors are the most common.

The extensive coverage of the Syrian conflict is in many ways a window on the future of conflict data. The future will be characterised by too much information and data of unclear quality. Internet access has increased reporting capacity and empowered citizens, as well as local organisations and belligerents to report. This has led to a proliferation of data, with some databases containing over 100,000 different events since the start of the war in Syria. But quality concerns abound. Many organisations are relatively new (a lifespan of 3 to 4 years is common) and do not all have a background in data collection – many have been learning on the job. As a result, there are visible data issues, the extent of which depends very much on the organisation. For example, key concepts are sometimes not clearly defined (thereby making data not comparable and sometimes unreliable). There are clear differences in methodologies, with some having them well developed while others have not. Moreover, organisations have specialised to such an extent that there is no initiative having a complete grip on all violence in the country.

Our partners

This project rests on collaboration with local partners in Syria. Many of them contributed under the expectation of anonymity. For this reason, some contributing organisations are anonymised in this report.

Data was generated from thirteen organisations. These thirteen were selected out of all initiatives reporting on the conflict (see Annex 1). Selection was based on: 1) whether they had clear, relatively defined concepts and methods; 2) whether they were transparent about their activities; 3) their geographical coverage (we collaborated with organisations covering large parts of the country); and 4) what sources they independently collected (we particularly probed for organisations who independently collected information in the country). Based on these criteria we approached various organisations, out of which one did not want to cooperate. Six organisations provided information that was not publicly available.

Table 1 Overview of the organisations involved in the pilot project

Contributing organisations		
1	Syrian Archive	Public
2	Carter Center	Private
3	Airwars	Private/Public
4	Liveuamap	Private/Public
5	Undisclosed source 1	Private
6	Syrian Human Rights Observatory (SOHR)	Public
7	Syria Direct	Private
8	Sham News	Public
9	Sana	Public
10	LSE-ISDC	Private
11	ISW	Public
12	Undisclosed source 2	Private
13	Standard Nexis Search	Public

Research design, method and description of the database

The research design involved two steps: a collection phase and a comparison phase.

Data collection

The project assessed the coverage of each organisation for three periods in the timespan from 2014 to 2016. It collected data on Week 49 of 2014 (December 1 to December 7), Week 5 of 2015 (January 26 to February 1) and Week 53 of 2015 (December 28 to January 3, 2016). In this way we ensured that we accounted for changing coverage over time (not all initiatives were covering each week). Subsequently, we realised that every organisation had different reporting metrics (e.g. data came in Excel files, verbal reports, videos and through visual images (photographs)). Therefore, we coded every initiative according to the existing ACLED methodology. All data was hand-coded into a common format to enable comparison (this format is described in Table 2 - Annex 2). Finally, information was coded by eight independent researchers hired for this purpose in the period from May to June 2017. This resulted in a database of 4,590 observations for the three weeks under consideration. These 4,590 observations were *not* unique events of violence but unique events *for each source.* Hence, an event could be present in SOHR data and in data from the Carter Center.

Data comparison

In the report we used two comparison methods. The first comparison used the database containing 4,590 events. Based on this data we assessed the amount of information covered, the geographical coverage, the coverage of actors and the type of violence initiatives reported. A second comparison metric was more advanced. We hired a research analyst to integrate all data. The analyst merged all data into one database (technically called matching).[7] This involved comparing data from two (or more) partners, determining whether these were similar events and keeping track of how much of a separate conflict event each covered. It led to a new database, without any duplicates, of 2,456 events (implying an average of about 800 distinct events a week).[8] The advantage of this metric was that it allowed an assessment of the uniqueness of each source. Throughout the report, the former database is called Database 1 (or duplicated data) while the latter is called Database 2 (or matched data).

Descriptive results

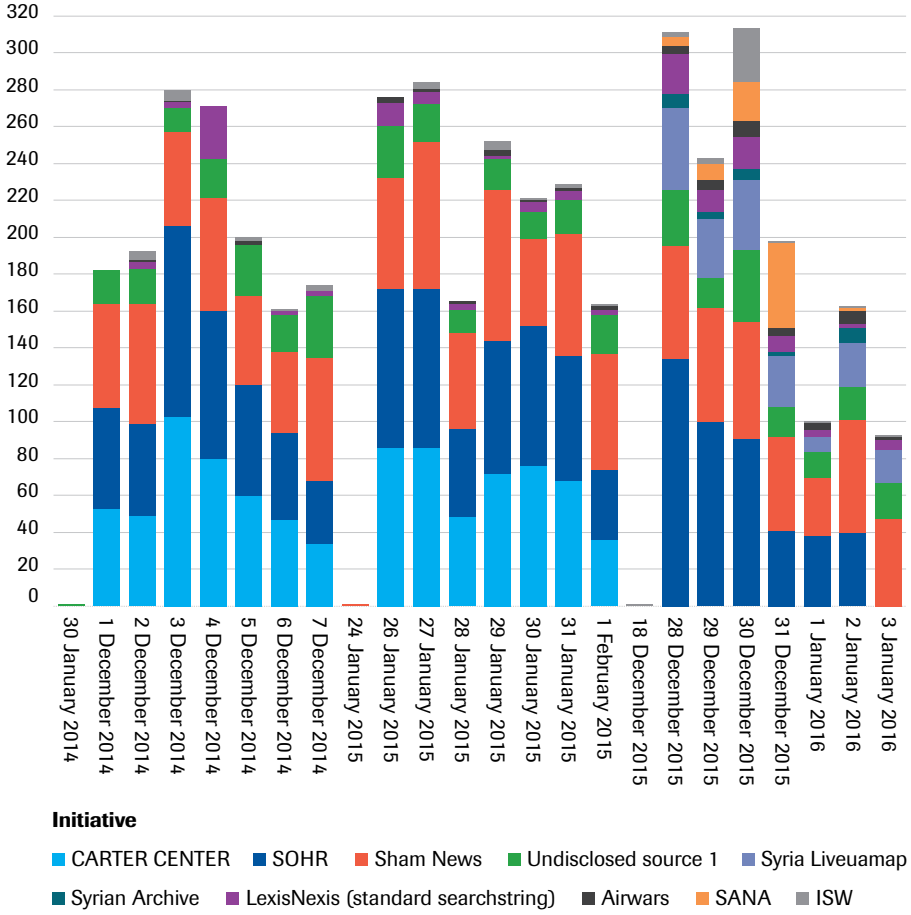
Descriptive results are contained in Figure 1. Each week of data is represented in a bar graph. The colours represent each initiative while the size of the bar indicates how many events the initiative contributed to. As is clear from the figure, there has been a variation in coverage by organisations. Some partners did not have information available for all of the three weeks – often because they started (systematic) coverage activity only around the end of 2015.[9] Some organisations provided information that was comparatively small (e.g. 7 events a week by one of the partners) and in the end we decided not to include these data. Moreover, we decided, due to the very high degree of overlap between the Carter Center and SOHR, not to code Carter data for the final week.

7 Xiaochen Zhu et al., "Matching Heterogeneous Event Data" (ACM Press, 2014), 1211–22, https://doi.org/10.1145/2588555.2588570; Stacie B. Dusetzina et al., *An Overview of Record Linkage Methods* (Agency for Healthcare Research and Quality (US), 2014), https://www.ncbi.nlm.nih.gov/books/NBK253312/.

8 See the conclusions on how matching could be used in other instances.

9 As is clear, the number of organisations covering the conflict increased over time (with 10 initiatives reporting in the last week of the project).

Figure 1 Coverage of the project



2 Coverage of the actors, locations and event types

Based on the collection of data on Syria we are able to present the findings on the different ways in which initiatives capture information on the Syrian conflict. While it is technical, this information is relevant for policy-makers for two reasons:

1. The identification of coverage patterns allows for a proof of concept. Syria's strategy for well-designed data can be replicated (see section 3); and

2. Varying coverage of sources underscores the need for the careful consumption of information on conflicts. For example, coverage from major Western reporting agencies narrate a different story than local sources such as Sham and even SOHR. Between local sources there are also major differences.

This section starts with general observations on data on Syria and subsequently moves to an assessment of reporting on: a) the geographical distribution of violence; b) the types of violence; c) the coverage of distinct actors. The main findings are summarised in the conclusion.

General observations

From Figure 1 it is already visible that the amount of data generated on Syria surpasses the capacity of individual organisations to cover the conflict. For example, the amount of data generated from three *weeks* of the Syrian conflict is comparable to over three *months* of data generated on the whole African continent.[10] Nearly all organisations confided to us how backlogs of data have only increased in 2017.

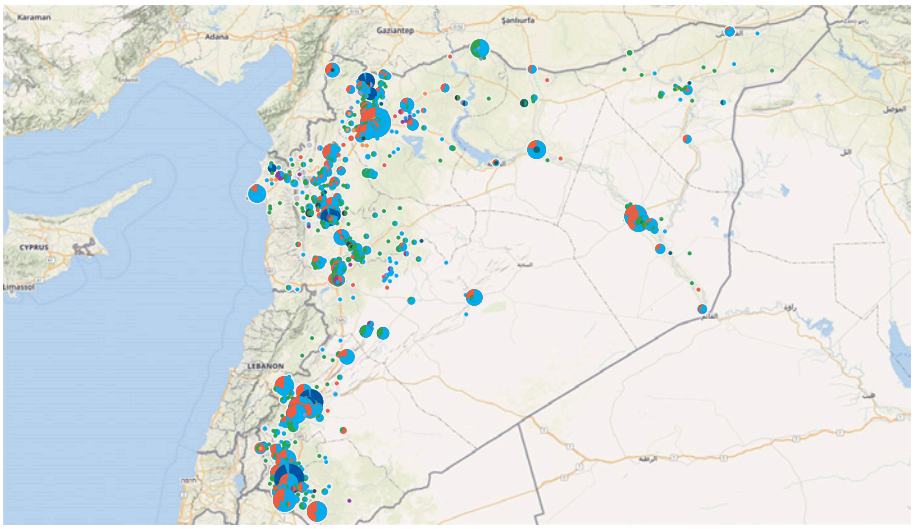
A second and third general observations are closely related. On the one hand, it is clear that four organisations provide the majority of events (SOHR, Sham, Undisclosed source 1 and Liveuamap). On the other hand, a third of all events (1,004 out of 2,456) are covered by more than one organisation. This implies that there is a duplication of efforts. The question therefore is what each initiative's unique contribution is and how it can be optimised. To assess this, we consecutively discuss geographic, event type and actors' coverage.

10 Based on ACLED data from 2013–2017.

Geographic coverage

Figure 2 presents the coverage of violence in Syria on a map. The primary conclusion is that all data combined the results in a comprehensive coverage of the conflict (for example, we have 770 unique locations in three weeks). Given that many events will have taken place in major cities (e.g. Aleppo, Damascus, Homs, Deir-ez-Zor and Raqqa), this coverage suggests a sufficiently detailed picture in almost all administrative areas. A second observation is that Raqqa and Al-Hasakeh are most likely not sufficiently covered.[11] Figure 3 (in Annex 2) shows that with 1.91 and 1.87% of all events respectively Raqqa and Al-Hasakeh are not sufficiently reported on. It is likely that the lower detection rates in these areas is a product of the networks of our partner organisations, with probably less developed networks in IS-controlled areas.[12]

Figure 2 Geographic distribution of events



11 Lattakia and Tarsus are known to have experienced less violence in the three-week pilot period.

12 Megan Price and Anita Gohdes, "Searching for Trends: Analyzing Patterns in Conflict Violence Data," *Political Violence at a Glance* (blog), April 2, 2014,

Aside from this good coverage, there are geographical reporting differences between our partners, with some reporting being better from particular areas. We use two metrics to show this: a) how well organisations are able to identify unique locations; and b) in which governorate they are generally able to do so. Firstly, regarding unique locations, Table 3 highlights that organisations capture varying degrees of unique locations (SOHR, Sham and Undisclosed source 1 capture many). The second metric shows that organisations have a different capacity to cover some districts (see Table 4 - Annex 2). For example, SOHR is better able to identify unique locations in Deir-ez-Zor and Damascus. Undisclosed source 1 identifies other locations in Aleppo and Sham and is better in Rural Damascus, Al-Hasakeh and Hama.

Table 3 Detection rate for new locations per initiative

	# Events	# Locations	# Unique	% Unique
SOHR	1,359	504	205	40.67
Sham	1,222	468	161	34.40
Undisclosed source 1	444	201	49	24.38
Carter	934	386	0	0.00
Nexis	151	85	16	18.82
Syria Live Map	232	76	15	19.74
ISW	82	54	13	24.07
SANA	83	63	13	20.63
Airwars	36	25	5	20.00
Syrian Archive	28	12	1	8.33

Coverage of types of violence

The methodology used distinguishes between nine types of violence or 'event types' (see Annex 2).[13] The results of the analysis lead to three conclusions.

A first conclusion is that the range of events we capture is roughly in line with the general violence profile in the Middle East. From a comparative angle and despite all the obvious limitations of our small sample, the data highlights that the conflict is more military (more remote violence and battles) than any of the other conflicts covered

13 Note that event types are **not mutually exclusive.** For example, a battle may involve civilian casualties but these are not coded as violence against civilians. Similarly, battles may include significant shelling (remote violence).

in ACLED data over the last 20 years. This snapshot is roughly comparable to other conflicts in the Middle East (compare, for example, Syria to Iraq, Table 5).

Table 5 Event type coverage compared

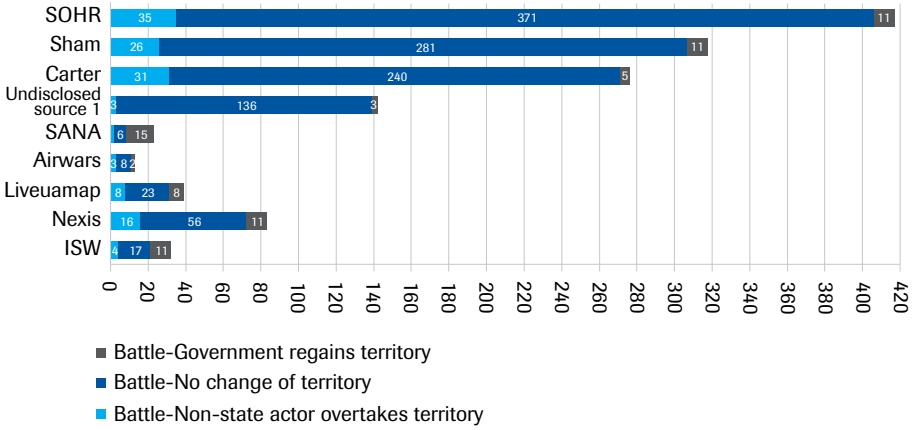
	Syria	Middle East	Iraq	Africa	Libya	Asia	Pakistan
Remote violence	60%	**34%**	49%	**6%**	22%	**3%**	8%
Battles	34%	**34%**	44%	**30%**	37%	**9%**	13%
VAC	4%	**8%**	5%	**28%**	16%	**6%**	5%
Riots/Protests	0%	**21%**	1%	**27%**	17%	**81%**	73%
Other	1%	**6%**	2%	**9%**	8%	**1%**	1%

A second conclusion is that we are likely to miss information on Violence Against Civilians (VAC). On the one hand, qualitative descriptions of the Syrian conflict tend to point to much higher levels of VAC defined as purposively targeting civilians rather than 'collateral'. Think, for example, of IS targeting civilians, killing in the prisons and recruitment drives. On the other hand, contributing partners have distinct reporting patterns on VAC. The largest contributors (the Syrian Observatory for Human Rights, the Carter Center and Sham) record far less purposive violence against civilians (1.5 to 3%) than an organisation like Undisclosed source 1 (23%). Since overlap measures are one of the only real ways to assess the comprehensiveness of coverage it seems there is more VAC than we can currently capture.[14]

A final conclusion is that there is a great deal of variation in military victories and that we are likely to understate the success of governmental victories. Opposition-'leaning' sources (SOHR, Sham and to some extent Undisclosed source 1) report fewer government victories while Sanaa - a government source – reports far more government takeovers. Both are probably not correct as other – more 'neutral' - sources report a higher number of government takeovers than opposition-leaning sources and a lower number than government-leaning sources (see Figure 4).

14 Cullen S. Hendrix and Idean Salehyan, "No News Is Good News: Mark and Recapture for Event Data When Reporting Probabilities Are Less Than One," *International Interactions* 41, no. 2 (March 2015): 392–406, https://doi.org/10.1080/03050629.2015.982117; Megan Price, Anita Gohdes, and Patrick Ball, "Technical Memo for Amnesty International Report on Deaths in Detention," *Amnesty International* 18 (2016).

Figure 4 Territorial takeover



Coverage of actors

A final piece of analysis pertains to the ways in which 'actors' are reported on by organisations. First, some organisations are much better at identifying unique actors. And second, the number of unidentified actors is too large.

First, some organisations are much better at identifying unique actors. Two organisations are particularly good. Undisclosed source 1 provides unique information on actors, having 28 actors not covered by any of the other initiatives. This is particularly impressive when considering the number of events (444) it needed to identify these 28 actors. SOHR, for example, added 18 unique actors and needed 1,359 events to get to that number (Liveuamap and a standard query in LexisNexis are also good). Sham is comparatively not doing well: it reports fewer actors and also not many new actors, which suggests that Sham's network is less diverse than SOHR's.

Second, there are some serious concerns about the amount of events where no organisation was able to verify the identity of the actor engaged in the violence. Unidentified actors make up between 20 and 30% of all reported incidents (see Figure 5 – Annex 2). Most of these events are airstrikes. Totally solving this problem will be impossible. However, some initiatives are twice as good at identifying actors (Undisclosed source 1, Nexis sources and ISW) than others (e.g. Sham).

Conclusion

The analysis of data collected leads to seven conclusions that are relevant for the ways in which policy-makers obtain, use and analyse information on Syria:

1. There is not one single comprehensive source on Syria. The bigger organisations do not cover more than 50% of all the data on their own. Policy-makers could support initiatives for genuine collaboration and the sharing of information;

2. Consumers of much of the information on Syria will usually obtain less information on incidents in Al-Hasakeh and Raqqa. There are specific strategies that are needed to increase information (e.g. the funding of organisations in those areas or specialised press to obtain reliable information);

3. Consumers of information on Syria will miss out on the specific targeting of civilians. These include forced recruitment drives, rape, mutilation and the targeted killing of civilians. Some specialised organisations (such as Undisclosed source 1) covered these incidents better than others;

4. The majority of the information stems from sources which lean towards the opposition. These sources tend to report a much larger number of opposition victories than is likely to be the case;

5. Consumers of credible local information from Syria are missing important information on belligerents. A number of the larger sources provide relatively general information on the actors (e.g. not reporting on the specific pro-government militia responsible for violence or only noting pro-government militias). For 20-30% of all events it is not clear who is responsible.

6. While not assessed in much detail here, additional analysis indicates well-known biases in the data. On the one hand, there is more reporting from populated places, reporting closer to road networks and reporting from urban areas. On the other hand, sources contained in LexisNexis report mainly on a war against jihadist organisations (especially on coalition strikes). Saana – a 'pro-government' news outlet – is consistently reporting a high number of government activity (often these are also government territorial takeovers). Sham – a 'pro-FSA' outfit – has a lower number of violence against civilians by FSA troops.[15]

15 Davenport and Ball, "Views to a Kill"; M. Herbert Danzger, "Validating Conflict Data," *American Sociological Review* 40, no. 5 (1975): 570–84, https://doi.org/10.2307/2094196.

7. Finally, there are serious quality concerns. Out of more than 30 initiatives collecting Syrian information, only half of them explained how this information was collected and what methods were used. Yet, even for those remaining 15 there are major quality differences.

All this leads to the need for a clever combination of sources that uses the strengths of partners to create a reliable and comprehensive picture. Therefore, the subsequent section details evidence-based steps and a 'proof of concept'. This concept can be used in other contexts as well.

3 Quality data by design

How can reliable data be produced? This section provides a step by step manual on the ingredients needed to create a reliable data set. Based on this project the fundamental idea is that three steps need to be followed to obtain reliable data sets. The following steps detail what is needed in the Syrian case.

1. Combining information requires a common vocabulary;
2. The varying quality of validation procedures suggests a metric measuring data reliability;
3. A specific sampling strategy suitable for individual conflicts.[16]

A common vocabulary

The first of any reliable combination of sources is – as the Syria case highlights – a common denominator. Various organisations have used different definitions to describe violence, often in line with the mission of their own organisation. However, on the whole it means that information is not comparable. For example, there needs to be agreement on definitions of when a territory is controlled and when it is not, or when an event constitutes an aerial bombardment and when it is recorded as the killing of civilians, or whether six individual killings are recorded as six separate events or as one event with six fatalities. This project uses a proven methodology (from ACLED) to allow different organisations to communicate but in other contexts other methods may be possible.

In these other contexts, policy-makers and researchers will receive and use databases that may tell (slightly) different stories. Hence, policy-makers – like researchers – may desire to merge different databases in order to obtain a more comprehensive picture. To do that a number of technical steps need to be taken (including a detailed assessment about what each source covers) but as this project highlights, a common and agreed upon vocabulary is needed.

16 See also Clionadh Raleigh, Matt Batten Carew, and Andrea Carboni, "Conflict Environments and Coverage", Report (The Hague: Clingendael Institute, 2018).

Event reliability

A second step is the recognition that different organisations have different procedures to validate the occurrence and details of events. These procedures are not necessarily of the same quality. For example, there are criticisms of the information provided by the Syrian Observatory for Human Rights (SOHR), among other things because specific details on how information is gathered in Syria remains unclear.[17] Similarly, many of the reports appearing on social media trackers (Liveuamap) are checked by an algorithm that – despite various attempts – may be subject to trolling attempts and disinformation campaigns from governments. Instead, organisations like Airwars, Undisclosed source 1 but also a small organisation like the Syrian Archive have diligent procedures to assess reliability (e.g. requiring visual evidence or oral testimonies of events). Hence, the construction of data should not be a garbage bin where each source is treated as if its reporting has equal reliability. To account for this, policy-makers could start including the reliability of individual events based on underlying sources.

For Syria we created the following metric (which can be used in other contexts as well). The higher the number, the more reliable the event is. The metric is based on two dimensions: 1) the number of sources; and 2) the quality of the procedures that partners have in place. Table 7 describes the metric. For Syria, a classification of our partners is contained in Annex 3 (annually updated).

Table 7 Event reliability score

Quality of source		Score	# of sources
Sources with some level of evidence (e.g. visual) but no corroboration of the evidence	One-source report	1	1 source
		2	> 2 sources
Source with due diligence (e.g. two sources or a certain method). There is a variation in this category with some reports having more credibility than others.	Credible reports	3	1 source
		4	> 2 sources (1 credible and 1 one-source)
		5	> 2 sources (2 or more credible sources)
Sources with open and transparent methodologies and independent verification procedures	Validated reports	6	1 or more sources

17 tom_taylor_20, "Syrian Observatory for Human Rights' Reliability 'has Been Found Wanting' – Coalition," *Grasswire* (blog), July 10, 2017, https://www.grasswire.com/2017/07/10/syrian-observatory-human-rights-reliability-found-wanting-coalition/; "The Syrian Observatory: The Inside Story," Al Akhbar English, accessed December 19, 2017, http://english.al-akhbar.com/content/syrian-observatory-inside-story.

A design strategy

The design of a database on a given conflict must be preceded by a detailed analysis of the coverage of sources and the black spots. *Hence, the procedure described below is only applicable to Syria-based analysis detailed above and cannot be extended to other contexts*. For Syria, the amount of data and information is so large that justifiable shortcuts need to be made in order to maximise coverage and quality (the design is open so that the availability of additional resources allows for additional coverage). The design of the Syria-database is a three-tier process (see Table 8).

1. **Using baseline data**. Two sources on Syria provide the most data – Sham and SOHR – and both of them include information that is unique. The analysis highlights that the most feasible strategy is to take one of the sources as the baseline and engage in a targeted strategy to add information on weak spots. Our analysis highlights that SOHR is the most viable source to serve as a baseline. This is a controversial choice: SOHR is not sufficiently transparent about the underlying sources and critique is levied that the source is deliberately not reporting on all events. At the same time, SOHR is undeniably the most comprehensive source as it has the highest number of unique locations, event types and actors. Compared to Sham, its reporting profile is more diverse.[18]

2. **Targeted enrichment to complement the baseline**. The analysis highlights that Liveuamap and Sham provide unique information not contained in SOHR. Liveuamap reports on unique event types and actors. Sham is better in unique locations. While Liveuamap can be captured within a reasonable amount of time, the amount of data generated by Sham is so large that we only include areas where Sham is most different from SOHR. These areas are: Hama, Hasakeh and Rural Damascus.[19] These pragmatic choices come at the cost of not increasing event reliability, as most events will obtain an average score of three (out of six). The availability of additional resources, either now or in the future, will allow for a further improvement of the database.

18 Using SOHR means that most events will have an initial event reliability of 3.

19 Aleppo and Dar'a would ideally be included as well when additional resources are available.

3. **Targeted enrichment for known deficiencies in the data**. The analysis showed three common deficiencies. One, the under-reporting of VAC. Second, the under-reporting of events in Al-Hasakeh and Raqqa and, finally, the unclear identity of the actors involved in remote violence (IEDs and airstrikes).[20] Based on this outcome we probed for additional open sources and approached specific organisations which are known to report on these deficiencies. After testing various organisations and twitter feeds we ended up with an additional set of sources.

Table 8 Targeted strategy for Syria

Deficiency		Source	Unique events
-	Baseline	SOHR	250 per week
0	Targeted increase	Liveuamap	40 per week
		Sham (Hama, Hasakeh, Rural)	60 per week
1	Violence Against Civilians (VAC)	Undisclosed source 1	75 per week
		Undisclosed source 2	30 per week
2	Coverage from Al-Hasakeh & Raqqa	RBSS	10 per week
		Hawar News Agency[a]	6 per week
3	Unidentified actors	Airwars (Coalition airstrikes)[b]	10 per week
		UNSC – SG reports (Airstrikes)	12 per week

a There is a very strong overlap with Euphrates Post suggesting that one of these sources is sufficient (and also that we are likely to be relatively comprehensive).
b A test of the inclusion of all ISW reports (detailing Russian/Israeli airstrikes) resulted in too few events.

Conclusion

The targeted strategy in Syria results in data that is better, more reliable and less biased than any of the databases currently in operation. At the same time, it should be understood that despite all of the efforts undertaken, biases will remain. There will remain under-reported acts of violence from rural areas and very small villages. Coverage will be impacted by targeted media blackouts, it will have a slight leaning towards the opposition and targeted violence against civilians is still likely to be under-represented (see Annex 3 for an overview). Moreover, given the inherent limitations in resources we will not be able to improve the reliability of individual events beyond otherwise very reasonable standards.

20 Almasdar News (Pro-government) may be a good source to balance the somewhat opposition-leaning nature of the data in the database. Addressing under-reporting from rural areas with no phone coverage may only be possible using time-consuming field methods, and finally the addition of information on headquarters and bases.

Be that as it may, the proof of concept presented in this report and the accompanying analysis provide the most well-researched attempt to gather and integrate reliable conflict data on Syria. It expressly targets areas where data is proven to be lacking and is a genuine collaborative effort open to new participants and organisations which strive for data quality. Our transparent and open system allows for the quality of conflict reporting to be increased as more information becomes available.

Annex 1 Reporting activity on Syria

Reporting activities

	Description	Sources	Time	Link
Syria Direct	Syria Direct is a journalism platform reporting on the Syria conflict from the perspective of Syrians. Reports on a wide range of activities including rebel to rebel violence. US Funded.	Sourcing through reporting network common journalism practices. No journalists in Syria, mostly in Jordan and Turkey.	2013-present	www
Sham News	Media outlet that aggregates photos and videos from citizen journalists in Syria. Activist news organisation critical of Assad regime.	Local sources through own reporting network of citizen journalists.	2012-present	www
Syria Deeply	News story-telling of conflict situations. Focus on political analyses, truces, attacks and troop movements.	Uses other media (UN News Center, Council on Foreign Relations, International Crisis Group, Human Rights Watch, BBC News – Syria Landing Page, Syria Comment, Syrian Revolution Digest, Syria Tracker, Now Lebanon, Global Voices. Currently mostly Reuters and AP).	2013-present	www

Human rights monitors

	Description/Goal	Sources	Time	Language	Link
Syrian Revolution Martyr Database	Records individual victims in Syria. Records age, gender and cause of death.	Combination of five sources (VDC, SHCR, Undisclosed source 1, Syrian Revolution database and Location Coordination Committees for Syria). Some additional records retrieved from social media and media.	2011-present	English/ Arabic	www
Syrian Centre for Statistics and Research	Records individuals who are dead, missing or have been arrested. By name, date and location.	Local network of reporters and a team of researchers inside and outside Syria.	2011-present	English/ Arabic	www
Raqqa is being slaughtered silently	Reports violations by IS and Syrian government in Raqqa and surroundings.	Formerly mostly activists in Raqqa, recent reports draw on contacts outside Raqqa.	2014-present	English	www
Syrian Tracker	Records violations such as killings, torture, massacres or rape. Categories changed and expanded over time without back coding.	Crowdsourced reporting, data mining from other websites (as a result, original sources are very hard to verify).	2011-present	English	www
Committee for the Defense of Democracy, Freedoms and Human Rights in Syria	Records individual deaths and casualties, including locations and names.	Claims it uses social media, but unclear exactly what types of sources and through which procedure.	2010-present	Arabic	www
Damascus Center for Human Rights Studies	Records individual casualties, and victims of extrajudicial killings, massacres, arbitrary detention, enforced disappearances, rape and torture (including names).	Social media reports verified through Syrian activist network (local network).	2012-2014*	English	www
Violation Documentation Centre (VDC)	Records missing, arrested and killed individuals including name, cause of death, location and actors involved.	In-country staff and contacts (local network).	2011-present	English	www

Syrian Human Rights Committee (SHCR)	Reports daily casualty counts and conflict events (shelling, VAC, clashes).	Unclear.	2011-present	English	www
Syrian Observatory for Human Rights (SOHR)	Records conflict events, stories, casualties, missing and detained.	Local sources, correspondents and activists (who may be informed through social media) ~ 200 reporters.	2011-present	English	www
Undisclosed source 1	Records detainees, deaths and attacks on vital facilities by one of the six main parties.	Network of local sources in Syria (▷ 1000). Each report is validated through testimonies and where possible photos.	2012-present	English	www
Syrian Archive	Records and archives human rights violations such as massacres, arbitrary arrests/detentions, torture, gender-based violence, illegal weapons, sieges, forced displacement and chemical attacks.	Various. Human rights organisations, Syrian research organisations, field hospitals, Local Councils/Coordination Committees, Local networks (activists, citizen journalists, lawyers), validated social media accounts. Collaboration with Bellingcat and Berkeley.	2012-present	English	www

* Whether Damascus Center is still operational could not be ascertained.

Conflict monitors

		Description	Sources	Time	Link
Carter Center	*Type*	Records 'conflict events' giving dates and exact locations. Events include bombings/ shelling, clashes, IEDs/suicide bombings and territorial takeovers.	Most information is coded from SHOR but Carter increasingly includes information from Facebook, YouTube, Twitter, Fora and other media.	2015-present	www
	Actors	Highly disaggregated (e.g. specific battalions and brigades of the various fighting forces).			
ISW	*Type*	Territorial control, airstrikes and weapons used. Occasionally other information such as troop movements or detailed city control maps.	Primary sources: ISW Syria team (local reporting network). Secondary sources: WikiMapia, SHOR, Syria Direct, Sham News Network. Sources changed over time. Grading of source reliability.	2013-present	www
	Actors	Assad regime, Hezbollah, JN, IS, FSA, YPG, Coalition (international), Allies.			
Airwars	*Type*	Recording of Coalition and Russian airstrikes. Rigorous validating of country responsible, location and casualty numbers.	Local sources, official media (international and local news agencies), local NGOs, social media (residents' Facebook, YouTube, twitter, fragmentary social media), Governments. Reliability classification.	2014-present	www
	Actors	US, UK, France, Netherlands, Australia, Denmark, Canada, Belgium, Russia.			
Liveuamap	*Type*	Territorial control. Maps with geo-plotted news-events.	Automated creation of maps from news reports and social media. Robot algorithms scrape data and interpret them.	2015-present	www
	Actors	Various: Regime, International Coalition, FSA (and associated groups), Kurds, Turkey, IS.			

Syrian Civil War Map	*Type*	Maps reporting on captured villages. Also maps with territorial control (used among others for WikiMapia/Wikipedia).	User-generated content and wisdom-of-the-crowd validation. Original sources are not reported on.	2015-present	www www
	Actors	SDF, IS, Opposition, Regime, Turkey.			
ACAPS/SNAP	*Type*	Various projects. Irregular provision of maps detailing territorial control and occasional coding of clashes.	Primarily relying on SHOR.	2012-2015	www
	Actors	FSA, YPG, SAF, Coalition (international), IS.			
INSO	*Type*	Security events of all kinds (thefts, threats but also assaults and attacks).	Information reported to INSO by its NGO members.	2014-present	www
	Actors	General: Assad Regime, Local Ethnic Militias, Coalition, OAG (organised armed groups).			
IHS Conflict Monitor (Jane's)	*Type*	Territorial control, no. of violent events, type of weapon used.	Local news sources, interpersonal intelligence gathering (HUMINT) and social media (300+ social media accounts validated for reliability and value).	2014-present	www
	Actors	Highly disaggregated (over a thousand armed factions in Syria).			
PAX Siege Watch	*Type*	Monitors besieged areas (using three levels) and areas under high risk of becoming under siege.	Reporting contacts on the ground, often affiliated with local councils. When unavailable through medical offices or citizen reporters.	2016-present	www
	Actors	FSA, NDF and various disaggregated actors.			

Other: ISDC-LSE (data, location/type violence), Crisesnet (unclear), Wikipedia (maps, villages taken), UCDP (maps, location violence), Cizire Canton (maps, territorial control)

Annex 2 Extra tables and graphs

Table 2 ACLED Event types

Event Types	Definitions
Battle - No change of territory	A battle between two violent armed groups where control of the contested location does not change.
Battle - Non-state actor takes over territory	A battle where non-state actors win control of location.
Battle - Government regains territory	A battle in which the government regains control of a location.
Headquarters or base established	A non-state group establishes a permanent or semi-permanent base or headquarters.
Strategic development	Activity by an armed actor that does not involve active fighting but is relevant for future violence. For example: recruitment drives, incursions, peace talks and arrests of high-ranking officials.
Riots/Protests	Non-violent (protest) or violent (riot) public demonstration by a group.
Violence against civilians (VAC)	Armed group attack on civilians. Civilians are unarmed and do not engage in violence.
Non-violent transfer of territory	Armed actors acquire control of a location without engaging in violence.
Remote violence	Activity where the tools for engaging in violence do not require the physical presence of the perpetrator (bombings, IED attacks, mortar attacks and missiles).

Figure 3 Coverage governorate

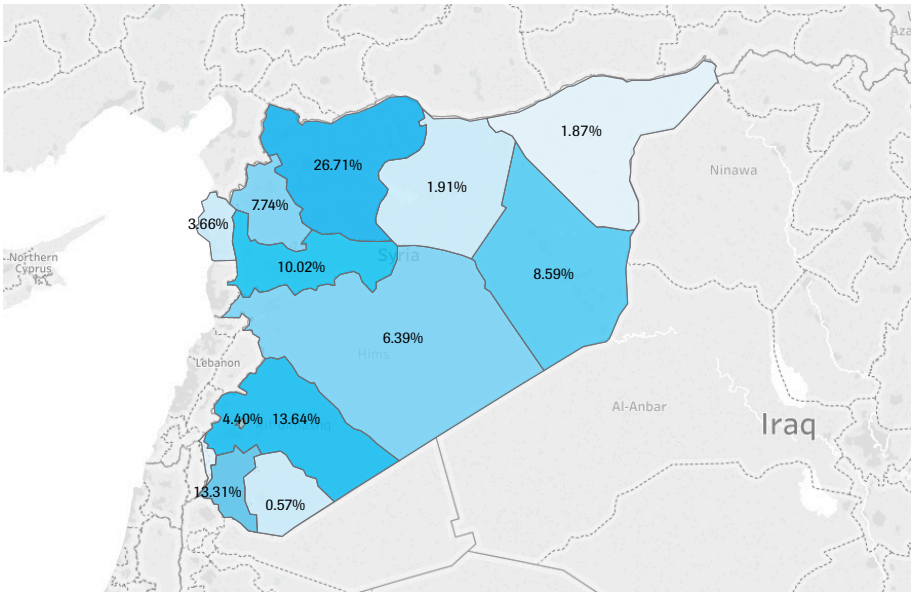


Figure 5 Identifiable vs. unidentifiable actors

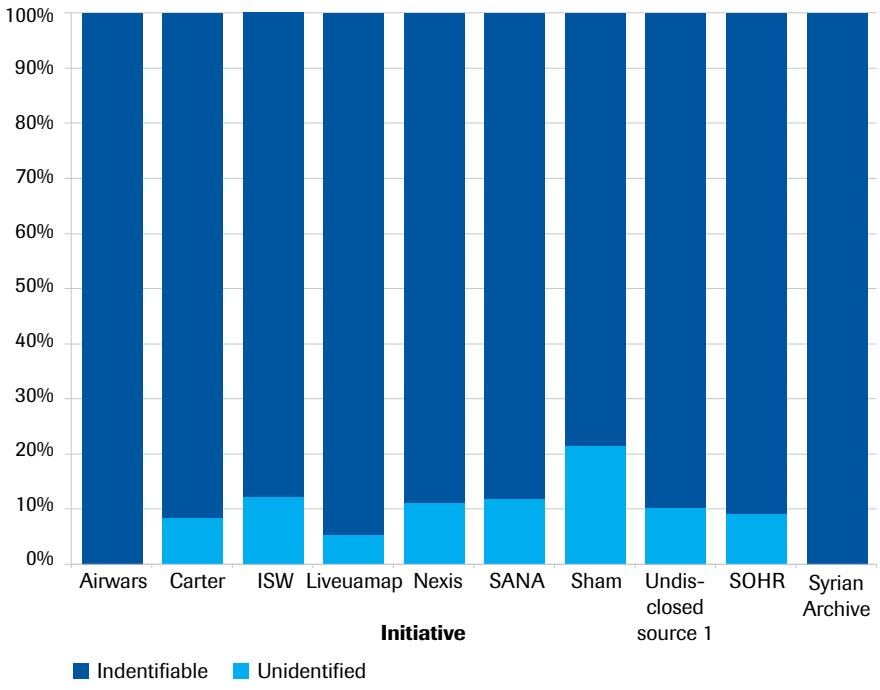


Table 4 Governorate coverage

		Sham	SOHR	Undisclosed source 1
Rural Damascus	**Unique**	37	28	8
	Not Unique	22	23	13
Al-Hasakah	**Unique**	18	4	3
	Not Unique			
Hama	**Unique**	15	11	5
	Not Unique	3	4	1
Aleppo	**Unique**	53	58	26
	Not Unique	16	21	6
Dar´a	**Unique**	45	29	10
	Not Unique	38	38	14
Damascus	**Unique**	15	13	9
	Not Unique	3	2	1
Deir-ez-Zor	**Unique**	32	35	6
	Not Unique	24	19	8
Homs	**Unique**	10	9	5
	Not Unique	5	3	2
Idleb	**Unique**	11	2	6
	Not Unique	4	3	2
Lattakia	**Unique**	8	3	1
	Not Unique			
Ar-Raqqa	**Unique**	1	1	1
	Not Unique	1	1	
Quneitra	**Unique**	1	1	1
	Not Unique			
As-Sweida	**Unique**		2	1
	Not Unique			

Figures in green show higher reporting by an organisation.

Annex 3 Reliability sheet Syria

Sources used: Carter Center, Airwars, Liveuamap, Undisclosed source 1, Syrian Human Rights Observatory (SOHR), Undisclosed source 2, Sham News, RBSS, Hawar News Agency, UNSC – SG reports. **Sources not-used**: Sana, ISW, Standard Nexis Search, Syria Direct, Syrian Archive.

Amount of information captured: unclear (though likely to be high).
Reliability: reasonable (average score between 3 and 4).

Likely bias	Addressed
Description of events follows the interest of organisations	Yes
Organisations depict actors differently	Yes
Under-reporting from rural areas	No
Under-reporting with no road network	No
Under-reporting when there is a great distance from the capital	Yes
Under-reporting when there is no cell-phone coverage	No
Under-reporting of low-impact and minor events	Partly
Under-reporting where the organisation has no network	Partly
Violence is better reported when there is a large population	Partly
Killings in high-impact events more often reported	Yes
Increasing violence often leads to decreasing numbers of witnesses	No
Local culture and personality traits lead to silent witnesses	No

Source Classification

Airwars	Validated	Multisource classification of reliability / Open and transparent public methodology
Syrian Archive	Validated	1) content acquisition and standardisation; 2) storing / Open and transparent public methodology
Undisclosed source 1	Validated	1) collection by recorders; 2) collect visual proof (2 sources); 3) database / cross-checking; 4) archiving Open and transparent public methodology
UNSC-SG	Validated	OHCRH and internal records information / UN information center + public methodology for OHCRH
CARTER	Credible	Combination of one credible source and social media / No public methodology available
Euphrates Post	Credible	Journalist principles / No public methodology available
Hawar News Agency	Credible	Journalist principles / No public methodology available
ISW	Credible	Multisource information / No public methodology available
LexisNexis	Credible	Journalist principles / No public methodology available
RBSS	Credible	Journalist principles / No public methodology available
SANA	Credible	Journalist principles / No public methodology available
Sham News	Credible	Journalist principles / No public methodology available
SOHR	Credible	Claimed usage of local sources / correspondents and activists (200) / No public methodology available
Syria Direct	Credible	Journalist principles / No public methodology available
Liveuamap	One-Source	Automated scanner of twitter based on self-learning algorithm / No public methodology available
Twitter	One-Source	Personal observation and hearsay / No public methodology available
Undisclosed source 2	One-Source	Internal reporting / No public methodology available