Reliable data on the Syrian conflict by design

Kars de Bruijne[1]
&
Clionadh Raleigh[2]

December 2017

***

This report presents the results of a pilot conducted from May to July 2017 measuring conflict data on the Syrian conflict. A number of local initiatives contributed data. The overall question the pilot sought to answer was: to what extent are local organizations producing timely, consistent and reliable data on the conflict in Syria? This report highlights strengths and weaknesses of various data initiatives. Descriptive results are covered in Part I (section 2 to 4). They suggest that common reporting patterns for event-data are present in the data. At the same time, the strengths of various initiatives are complementary. Part II presents a proposal for the combination of data from different organizations. The proposal includes the usage of 'baseline' data and a two-step enrichment process to expand actor, location and event-type coverage (step 1) and a targeted query to improve coverage of specific details known to be under-reported (step 2). The report closes with organizational suggestions for the (meta-)database.

[1] Senior Researcher, Armed Conflict Location & Event Data Project (ACLED)/Research Manager Syria Project., Research Fellow Clingendael Institute.

[2] Professor of Human Geography, University of Sussex., Director, Armed Conflict Location & Event Data Project (ACLED).

Introduction

Conflict data is prone to substantial bias. For one, conflict researchers have little insight into the exact amount of all conflict-event that their databases capture.[3] Moreover, conflict data reproduces a number of the biases inherent to qualitative reporting. For example, journalists tend to stay within cities and, therefore, report less on rural areas that are harder to access[4]. Mobile phone coverage can improve reporting[5], but is not available everywhere. And finally, the networks of media and local organizations are influenced by their allegiance to parties in conflict.[6] These biases greatly affect conflict databases and, subsequently, analysis and trends produce which therefore may not truly reflect reality.[7]

The Syrian conflict is perhaps one of the best monitored conflicts of all time. A survey of a large number of organizations reporting on the conflict identified over 30 distinct initiatives each collecting information on violence in Syria. This project builds on the premise that a combination of many of these sources is the best way to build reliable data on Syria. When we have a good insight into what each organizations covers, we are well placed to correct for known biases as much as possible. To this end, this project tests the reliability, consistency and usability of a diverse set of data generated by various organizations in the country. It asks to what extent these organizations offer unique data on the Syrian conflict. Moreover, it probes how this data can be combined into a comprehensive database.

The project is carried out by a consortium of four organizations. The consortium is a collaboration led by the Armed Conflict Location Event Dataset (ACLED) and includes the Clingendael institute, International Security and Development Centre (ISDC) and the London School of Economics (LSE). This pilot project is a co-product of ACLED and Clingendael. They hired eight research analysists to code the data. Subsequently, two (senior) researchers analyzed the data. The results of the pilot and the proposed design for combining data on Syria is funded by the US State Departments' Conflict and Stabilization Operations (CSO).[8] The results/data will be made publicly available, free of charge.

To our knowledge this project provides the most comprehensive initiative to date of conflict data on Syria.[9] This report is for those organizations who contributed data and information to the pilot.

[3] Kars de Bruijne and Erwin Van Veen, "Pride and Prejudice: Recognizing Bias for Policy-Makers" (Den Haag: 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.

[4] 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.

[5] 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, 1065912916670272, https://doi.org/10.1177/1065912916670272.

[6] 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.

[7] We define bias as the difference between actual reality and reality as reported (or perceived). 'Actual reality' represents an individual or collective act of political violence that has occurred. Subsequently, these acts are 'constructed', for example in how they are reported, analysed or framed by perpetrators, researchers and journalists. Bias, therefore, occurs when there is a difference between 'actual acts of political violence' and 'constructed acts of political violence'.

[8] A substantial start has been made – see below.

[9] 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.

Research Questions

1. How do different data sources on the Syrian conflict differ with regards to the a) geographical distribution of violence; b) types of violence captured; and c) coverage of distinct actors?

2. How can these disparate sources be combined to create a reliable and consistent conflict database on the Syrian conflict?

This report begins by presenting the pilot's methodological choices such as the selection of the thirteen local partners, the Syrian context and research design and process. Following this, Part 1 presents the results of the pilot. It explains which organizations cover what locations, types of actor and types of violence, as well as 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 which draws upon them all. Part 2 subsequently, explains how the data can be combined into one reliable set of data and creates this data.

1. Motivation, partners and methods

The Syrian conflict is perhaps one of the best monitored conflict of all times. This extensive coverage of the conflict, is in many ways a window into the future of conflict data. The future of conflict data will be characterized by the existence of too much information to gather and which is of unclear quality. Internet access has increased reporting capacity, empowered citizens, local organizations and belligerents to report. This had led to a proliferation of data, with some database containing over 100,000 distinct events since the start of the war in Syria. But, with increases in quantity of conflict data, quality concerns abound.

Many organizations are relatively new (a life-span of 3 to 4 years is common) and not all have a background in data collection. Many have been learning on the job. As a result, there are sometimes issues with the data collected. For example, key concepts are sometimes not clearly defined (making data incomparable and sometimes unreliable). There are also differences in the maturity of methodology, with some organization having well-developed and clear methodologies while others lack methodological clarity. Moreover, organizations have narrowed their data collection focus, concentrating on particular types of violence or facets of the conflict, meaning that there is no one initiative which provides a complete picture of violence in the country.

Our partners

This project rests on collaboration with local partners in Syria. Data was generated from thirteen organizations (see Table 1). These thirteen were selected out of all initiatives reporting on the conflict (see Annex 1). Selection was based on: 1) whether they had clearly relatively defined concepts and methods; 2) whether they were transparent about their activities; 3) their geographical coverage (we collaborated with organizations covering large parts of the country) and; 4) what sources they independently collected (organizations who independently collected information in the country were sought out). Based on these criteria we identified and approached 14 organizations, one of which did not want to cooperate. Six organizations provided information that was not publicly available.

Table 1: Overview organizations involved in the pilot

Contributing organizations		
1	Syrian Archive	Public
2	Carter Centre	Private
3	Airwars	Private/ Public
4	Liveuamap	Private
5	Syrian Network for Human Rights (SNHR)	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 1	Private
13	LexisNexis	Public

Research design, method and description of database

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

Data collection

To begin, the coverage of each organization was assessed for three periods from 2014 to 2016. The three periods selected were Week 49 of 2014 (December 1– December 7), Week 5 of 2015 (January 26 to February 1) and week 53 for 2015 (December 28 to January 3 2016). The periods were (quasi-randomly) selected based on variation on coverage by potential partners. Several periods, spread out over time, were selected to ensure that changes in coverage over time were accounted for (not all initiatives were covering every week). Every organization had different reporting metrics (e.g. data came in excel files, verbal reports, videos and photos). Therefore, we coded the data of every initiative according to an existing methodology (ACLED's methodology). All data was hand-coded into a common format to allow for comparison. The coding was performed by eight independent researchers hired expressly to complete this task 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, for example, an event could be present in SOHR data and in data from the Carter Centre.

Data comparison

To compare the data, we used two methods. The first method used the 4,590 events database. We assessed the volume of data, geographical coverage, coverage of actors and types of violence reported. The second comparison method was more advanced. We hired a research analyst to integrate all data. The analyst merged all data into one database (technically called "matching").[10] This involved comparing data from two (or more) partners, determining whether the events they recorded were the same events or similar (duplicates), and then isolating the unique events reported by each initiative. This led to a new database without any duplicates which had 2,456 events (implying an average of about 800 distinct events a week).[11] This method allowed for an assessment of the uniqueness of each source.

Part I: Reporting on Syria

This part presents results from the pilot. The part starts with descriptive results and subsequently assesses how different data sources on the Syrian conflict report on: a) geographical distribution of violence; b) types of violence captured; and c) coverage of distinct actors?

Figure 1: Coverage Pilot

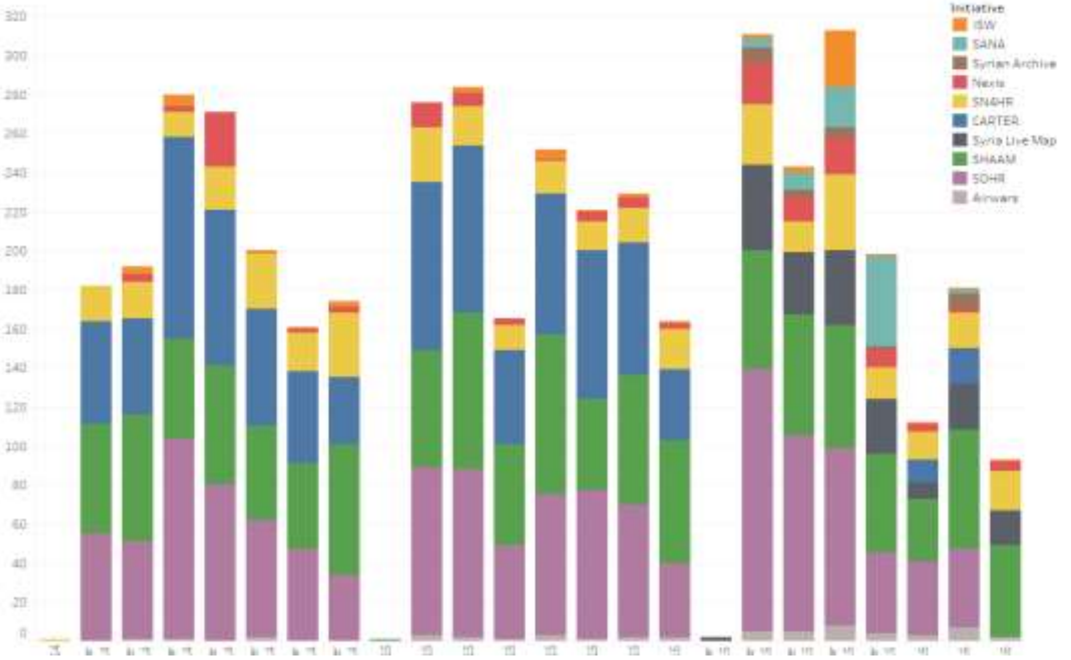


Figure 1 presents a descriptive overview of the data. Each week of data from the participating local data initiatives is represented in a bar-graph. The colors represent each initiative while the size of the bar indicates how many events the initiative contributed. As

[10] 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/.

[11] See conclusions on how matching could be used in other instances.

is clear from the figure, there has been variation in coverage by organizations. Some partners did not have information available for all three weeks – often because they started (systematic) coverage activity only around the end of 2015.[12] Some organizations provided information that was small comparatively (e.g. 7 events a week by one of the partners). After assessing the amount of data generated for one week, we decided not to include the data (given that the costs of sifting through all evidence did not weight to added coverage). Moreover, we decided that, due to the very high overlap between the Carter Centre and SOHR, we would not code the Carter Centre data for the final week. Overall, the following conclusions emerged from the data coding and comparison processes:

1. **The amount of data generated in Syria surpasses the capacity of individual organizations to cover the conflict**. For example, the amount of data generated from three *weeks* of the Syrian conflict is comparable with over three *months* of data generated on the whole African continent.[13] Nearly all consortium partners have highlighted how backlogs of data have only increased over the past year. This implies that genuine collaboration and (potentially) specialization is the most viable strategy to gain a good insight into the Syrian conflict;

2. **Coverage of the Syrian conflict varies over time. This implies that the consortium needs a regular scan of activities by new organizations**. Comparing the first two weeks (from 2014 and 2015) with the third week (end of 2015/beginning of 2016) highlights that coverage has increased and that more organizations have become active. For example, the social media scraper "Syrian Liveuamap" has reported activity from the middle of 2015 onwards and consequently is only represented in the last pilot week. Overall, this implies that a process needs to be put in place in order to identify new activity and coverage on a regular basis;

3. **Four organizations are responsible for the majority of events but their coverage often overlaps (see Figure 1).** The Syrian Observatory for Human Rights, the Carter Centre and Sham news each cover roughly 400+ events per week. The Syrian Network for Human Rights also reported a considerable number of events per week (about 150). Yet, rather than 1350 (3*400+150) they generated about 700 unique events.

Given these conclusions, a question emerged: what is each initiative's unique contribution with regards to the: 1) geographic coverage; 2) type of events captured and; 3) coverage of unique actors. Each element is discussed in the following sections.

2. Geographic coverage: standard biases are visible

The organizations in the pilot consortium identified events in Syria including their precise locations and specific geo-coordinates (latitudes and longitudes). This allows the data to be

[12] The amount of organizations covering the conflict increased over time (with 10 initiatives reporting in the last week of the project).

[13] Based on ACLED data from 2013-2017.

mapped and used to track developments over time. Based on the geographic coding of information the following four conclusions can be drawn:

1. **The consortium is comprehensive in covering unique locations of violence in Syria**. Out of 2,456 unique events covered in the pilot period, we record about 770 unique locations. With a high number of events happening in the same locations (Aleppo, Damascus, Deir-ez-Zor and Raqqa) coverage seems to be comprehensive and sufficiently detailed in almost all administrative areas. This is illustrated by Figure 2 which presents the data on a map. As is clear from the visual, there is a great geographic dispersion of events throughout the country.

2. **There are clear geographical reporting differences between the contributing organizations**. As Table 2 highlights, the number of unique locations captured by each organization varied considerably. For example, the SNHR covers 49 locations that are not reported by any of the nine other organizations (about 25% of all its locations are unique). Three organizations (SOHR, Sham and SNHR) are responsible for almost 90% of all unique locations. Another difference between the organizations is that some are better in reporting from certain areas. Sham and SNHR, for example, have a better coverage of Raqqa and Al-Hasakeh than most other organizations. This means that that the consortium needs a combination of sources to enable comprehensive coverage. Taking into account that not all initiatives can be covered, a viable strategy is taking the highest number of unique locations. This suggests using SOHR, Sham and SNHR. A third observation is that the main contributors of unique locations – SOHR and Sham – are good in different areas. As Table 3 (Annex 2) highlights, Sham reports particularly well on Hama and Al-Hasakeh (and to a lesser extent Rural Damascus and Dar'a) while SOHR is better on the other areas.

Table 2: Detection rate new locations per initiative

	# Events	# Locations	# Unique	% Unique
SOHR	1.359	504	205	40.67
SHAM	1.222	468	161	34.40
SNHR	444	201	49	24.38
Carter	934	386	0	0.00
Lexisnexis	151	85	16	18.82
Syria Liveuamap	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

3. **The data reproduces a number of well-known ´biases´.** There is more reporting from populated places, as well as reporting closer to road networks and reporting

from urban areas. This corresponds with known ´biases´. Conflict events are better captured when occurring in high population areas (hence cities rather than rural

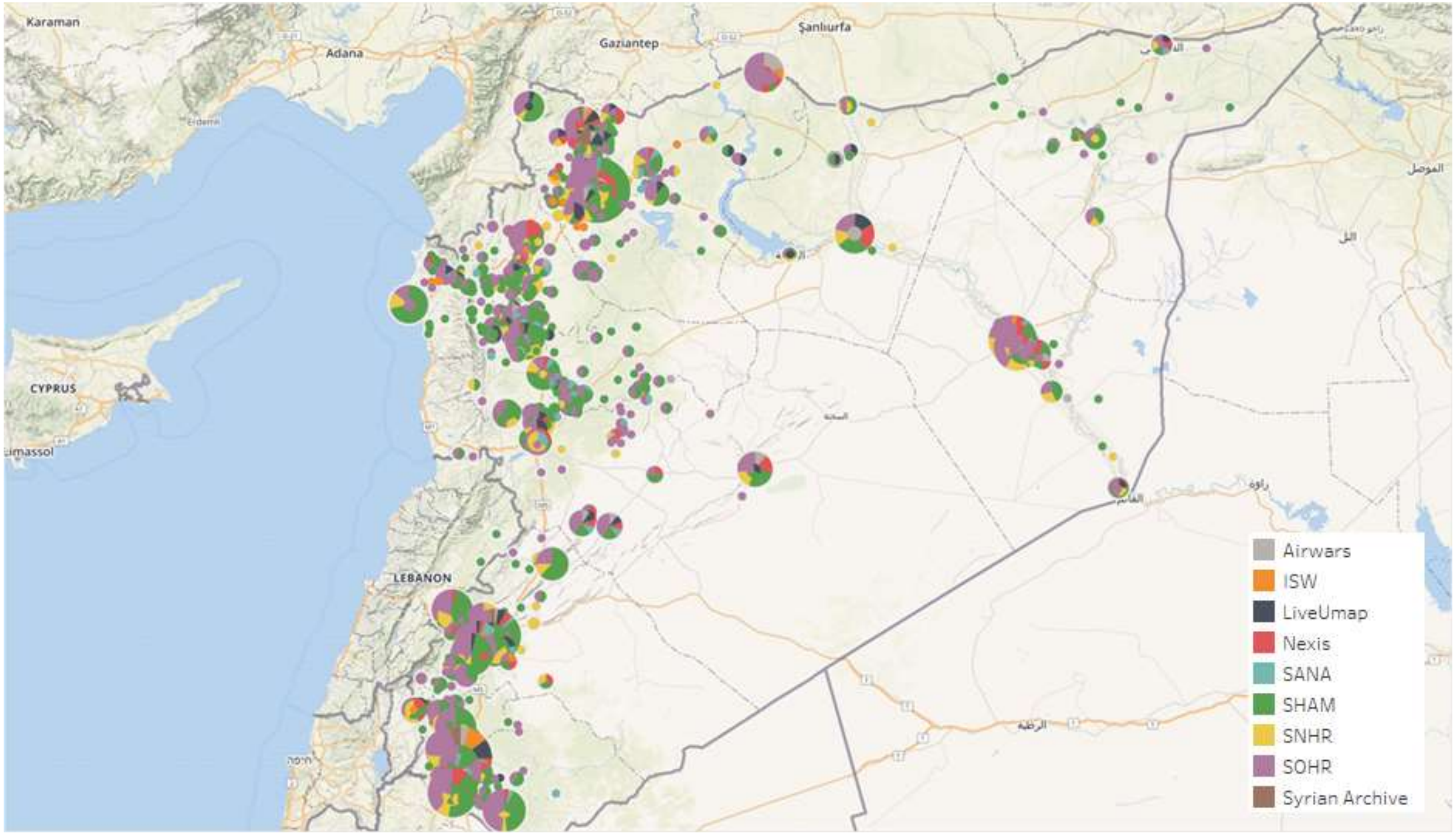


Figure 2: Geographic distribution of events

areas), closer to the road network and when they involve casualties (cellphone coverage may also be a determinant).[14] Recognizing these biases is to begin to correct for them;

4. **The consortium may be underreporting the districts Raqqa, Al-Hasakeh and, to a lesser extent, Homs** (Lattakia and Tartsus are known to have experienced less violence in the three-week pilot period). With 1.91% and 1.87% respectively, the coverage of Raqqa and Al-Hasakeh suggests underreporting (Figure 3 – Annex 2). It is likely that the lower detection rates in these areas corresponds to known biases/problems for disaggregated conflict data, but is also a product of the goals and networks of organizations.[15] Given that many organizations may have links to 'the opposition' it seems likely that there is underreporting from IS and government held areas. Hence, despite the enormous efforts of all organizations involved, there remain challenges in covering some areas that the consortium needs to overcome.

3. Even-type coverage: to little reported Violence Against Civilians

ACLED distinguishes between nine types of violence or 'event-types' (see Table 4). The consortium uses ACLED´s event-type categorization which is well-established in both academic and policy circles.[16] The main advantage is that the typology enables comparison of the Syrian conflict with other conflicts. 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).

Table 4: 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 overtakes 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

[14] 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; Matthew A. Baum and Yuri M. Zhukov, "How Selective Reporting Shapes Inferences about Conflict," 10040.

[15] Megan Price and Anita Gohdes, "Searching for Trends: Analyzing Patterns in Conflict Violence Data," *Political Violence at a Glance* (blog), April 2, 2014, http://politicalviolenceataglance.org/2014/04/02/searching-for-trends-analyzing-patterns-in-conflict-violence-data/.

[16] We are open to additions or changes when partners feel we are missing important dimensions.

	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 tool for engaging in violence do not require the physical presence of the perpetrator (bombings, IED attacks, mortar and missiles).

Using these event-types the consortium assessed the reporting profiles of organizations and the general 'reliability' of event-type reporting. Table 5 and Figure 4 (both in Annex 2) present descriptive statistics for the types of events captured by each individual organization. Table 6 compares the average distribution over event-types from Asia, Africa and the Middle East.[17] The following conclusions emerge:

1. From a comparative angle, and despite all the obvious limitations of our small sample, the data highlights that the conflict is more military (i.e. more remote violence and battles) than any of the other conflicts covered in ACLED data over the last 20 years. It is likely that the snapshot is roughly correct as conflict characteristics are comparable to other conflicts in the Middle East (compare for example Syria to Iraq (Table 6)). At the same time, the distribution of event-types also suggests that the consortium is likely to miss out on some important information: **violence against civilians and riots/protests appear under-sampled**. There are two reasons for this. First, qualitative descriptions of the Syrian conflict tend to point to much higher levels of Violence Against Civilians (VAC) defined as purposively targeting of civilians rather than as 'collateral damage'. For example, IS targeting civilians, killing in the prisons and violent recruitment drives. Rioting and protesting by civilians for and against warring groups is also relatively common but not often reported;

2. A second reason to believe that VAC and riots/protests are under-sampled is that **contributing partners have distinct reporting patterns on Violence Against Civilians**. The largest contributors (the Syrian Observatory for Human Rights, Carter Centre and Sham News) record far less purposive violence against civilians than an organization like the Syrian Network for Human Rights. It seems very likely that there is more VAC than currently reported by contributing organizations. On the one hand, SNHR's transparent and diligent methodology makes us confident that the events it reports are very likely to be correct. On the other hand, the only way to assess whether reporting approximates 'actual reality' (hence is unbiased) is to measure the overlap between organizations. The larger the overlap the more likely that the data is unbiased. However, the overlap for VAC in different governorates and between different organizations is low. For this reason, it seems very likely that there is more VAC than currently reported by contributing organizations;

[17] Please note that analyses are rough, since data has not been reduced to unique coverage.

3. Apart from an apparent underreporting of VAC and riots/protests there are **two additional minor biases on event-types**. First, there are different reporting characteristics on military victories. Opposition 'leaning' sources (perhaps SOHR, SNHR and Sham) report fewer government victories while SANA - a government source – reports far more government takeovers (see Figure 4, Annex 2). Both differ from sources where leanings to either party are less pronounced, either due to editorial constraints (lexisnexis, ISW) or due to the lack of a filter (social media). These sources report a higher number of government take-overs than opposition leaning sources and lower number than government leaning sources. A second minor bias is that none of these sources reports on the presence or establishment of headquarters. A specific query for these event types is therefore needed.

Table 6: Event-type coverage compared

	Syria	**Middle East**	Iraq	**Africa**	Libya	**Asia**	Pakistan
Remote violence	59.79%	**31.43%**	48.84%	**6.05%**	22.41%	**3.18%**	7.70%
Battles	34.44%	**34.42%**	43.52%	**30.37%**	37.33%	**8.72%**	12.79%
VAC	4.32%	**7.84%**	4.98%	**27.80%**	15.81%	**5.76%**	5.30%
Riots/Protests	0.26%	**20.52%**	0.81%	**26.70%**	16.81%	**81.03%**	72.96%
Other	1.19%	**5.68%**	1.85%	**9.08%**	7.64%	**1.27%**	1.25%

4. Coverage of actors: detecting well-known patterns

The design of the pilot included the identification of two actors for every event (except for one-sided events). In total, more than one hundred distinct actors were identified for the sample period. Given the three week coverage this is a reasonably high number, but probably a little too low given the highly fragmented conflict environment.[18]

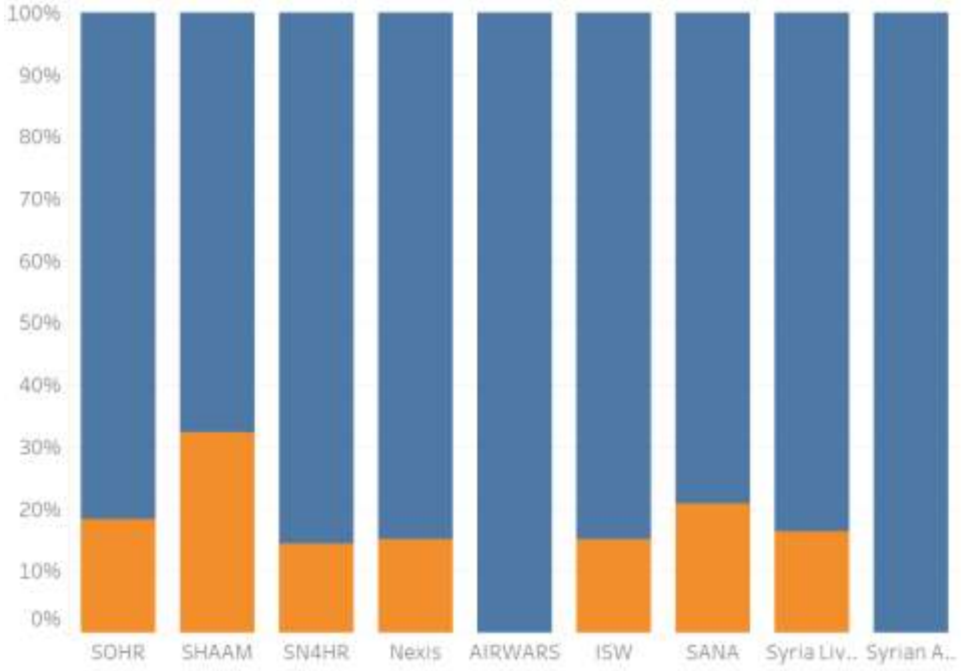
This section analyses how well organizations report on actors and also whether they were similar in their reporting on them. To assess this, we used as a benchmark the extent to which reporting was similar: the more similar reporting across sources the more likely the sample approximated the real number of events. Based on this comparison three conclusions can be drawn:

1. The consortium used a label for unidentified armed groups meaning that the identity of the actor could not be ascertained. Unidentified actors make up between the 20 and 30% of reported incidents in the pilot phases. Hence, our first conclusion is that **the number of unidentified armed actors is too high**. Given 600 events on a weekly basis we are unable to detect actor culpability (or targets) for up to 120 events per week (20%). This number is highest when the actors are involved in battles but is also high when detecting remote violence. Solving this

[18] What actors are reported on is, however, open and coverage will expand whenever a report identifies an actor.

problem is not easy, but strategies to improve identification are possible. We found that some initiatives are generally twice as good at identifying actors (SNHR, Nexis sources and ISW) as others (e.g. Sham) (see Figure 5). Moreover, remote violence incidents – particularly bombardments – greatly benefit from the work by ISW (on Russian and Israeli airstrikes) and Airwars (on coalition airstrikes).

Figure 5: identifiable (blue) vs. unidentifiable (orange) actors



2. A second conclusion is that **some organizations are much better in identifying unique actors**. Table 7 provides descriptive statistics showing a wide percentage difference of unique actors being covered. Two organizations are particularly good. SNHR provides unique information on actors having 28 actors that are not covered by any of the other initiative. This is particularly impressive when considering the number of events (444) it needed to identified these 28 actors. SOHR for example included 18 unique actors and needed 1359 events to get to that number. Syria Liveuamap and a standard query in Lexisnexis also performed well in this regard. Comparatively, Sham did not perform well in this regard: it reports fewer actors generally, as well as fewer new actors, suggesting Sham's network is less diverse than SOHR's.

	# of events	Not unique	Unique	% New
SNHR	444	22	28	56.00
SOHR	1359	28	18	39.13
AIRWARS	36	4	1	20.00
Liveuamap	232	17	4	19.05
Nexis	151	27	5	15.63
ISW	82	16	2	11.11
SHAM	1222	29	5	14.71

SANA	83	10	1	9.09
Syrian Archive	28	2	0	0.00

Table 7: Detection rate of new actors per initiative

3. A third conclusion is that **there are well-known and to-be-expected reporting biases in the data**. Such biases are often driven by both reporting and description biases in the sources.[19] Three examples attest to this. First, sources contained in Lexisnexis report mainly on a war against jihadist organizations with a high incidence of coalition strikes against IS. This suggest that most sources are Western based and appear to provide information domestic audiences are believed to be interested in. Second, Sana – a 'pro-government' news outlet – is consistently reporting a high number of government activity (often these are also government territorial take-overs), apparently in an attempt to please its pro-government readers. Finally, Sham news – a 'pro-Free Syrian Army (FSA)' outfit – has a lower number of violence against civilians by FSA troops. While these reporting characteristics are nothing new, they underscore the need to use a broad and diverse set of sources for reporting.

Part II: Quality data by design

Perhaps the most important conclusion on the basis of the pilot is that **not all available information can, nor should, be used.** There are two reasons. First, uncritically combining many sources is likely to lead to increased biases rather than an improved data-sample.[20] Second, the amount of information on the Syrian conflict is so large that the available resources and time do not allow for a complete inclusion of all possible sources. Therefore, the database-design is motivated by three considerations:

1. ACLED provides data in near real-time to enable informed evidence-based analysis;
2. This data maximizes coverage but also reliability and account for bias;
3. The databases should allow for *ex post* addition of information.[21]

Protecting contributing organizations is crucial. The consortium guarantees source protection and when desired will anonymize contributing organizations. Hence, information can publicly neither be traced back to organizations nor to individuals reporting to contributing organizations. Below we described our basic design (A), event reliability (B), and the step-by-step collection process (C).

A. Basic design

[19] 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; M. Herbert Danzger, "Validating Conflict Data," *American Sociological Review* 40, no. 5 (1975): 570–84, https://doi.org/10.2307/2094196.

[20] Taylor B. Seybolt, Jay D. Aronson, and Baruch Fischhoff, eds., *Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict* (Oxford University Press, 2013), https://doi.org/10.1093/acprof:oso/9780199977307.001.0001.

[21] The last requirement is particularly important for initiatives tracking civilian casualties. Many are faced with severe backlogs or rely on information that is only available after some time.

To merge databases, a common and agreed upon vocabulary on what constitutes an event is crucial. In Syria, various organizations have used different definitions to describe violence, often in line with the mission of the organization itself. However, as a whole this means that information is not comparable. For example, there needs to be agreement on definitions for when a territory is controlled and when it is not, or when an event constitutes an aerial bombardment and when it is recorded as killing of civilians, or whether six individual killings are recorded as six separate events or as one event with six fatalities.

The consortium has used ACLED's infrastructure to record events and proposes to continue with this set-up. The pilot has resulted in an expansion of coding rules and specific guidelines on how to deal with border cases. The pilot suggested no clear need to expand the ACLED-infrastructure and methodology. However, slight changes were made to account for the Syria case (for example the category strategic development has been expanded, civilians are now included as an associated actor when they are 'collateral damage' in remote violence incidents and bombardments are discerned from shelling). However, the consortium remains open to suggestions to improve the design.

B. Event reliability

A second step is recognition that different organizations have different procedures to validate the occurrence and details of events. These procedures are not necessarily of the same quality. For example, there is criticism of the information provided by the Syrian Observatory for Human Rights (SOHR), among others, because specific details on how information is gathered in Syria remains unclear[22] 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, organizations like Airwars, SNHR but also small organizations like the Syrian Archive have diligent procedures to assess reliability (e.g. requiring visual evidence or oral testimonies of events). Hence, the process of collection and compilation of data should not assume each source is equally reliable.

For Syria we created the following metric. The higher the number the more reliable the event is. The metric is based on two dimensions: 1) the number of sources reporting the event and; 2) the quality of procedures the source has in place to verify events. Table 8 describes the metric. For Syria, a classification of our partners is contained in Annex 3 (annually updated).

Table 8: Event reliability score

Quality of Source	Quality of Source	Score	# of sources
Sources with some level of evidence (e.g. visual) but no corroboration of	One-source report	1	1 source
		2	> 2 source

[22] 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.

the evidence			
Sources with due diligence (e.g. two sources or some method of verification). There is variation in this category with some reports being more credible than others.	Credible reports	3	1 source
		4	> 2 source (1 credible and 1 one-source)
		5	> 2 source (2 or more credible sources)
Sources with open and transparent methodologies and independent verification procedures	Validated reports	6	1 or more sources

C. Step-by-step collection process: from baseline to purposive enrichment

The outcomes of the pilot taught us that the amount of data and information is very large. To maximize coverage and ensure quality data within the constraints of time and resources, we propose a three-tier process that makes some justifiable shortcuts. The process is open so that at every point time new information can be added so that the quality of coverage of the database can expand. Table 9 provides the three steps we take to arrive at high-quality data. This involves the usage of ‘baseline data’; a targeted enrichment of the baseline data to engage in a general increase of coverage and reliability and; a targeted strategy aimed at identified weak spots in the coverage of sources.

1. **Using baseline data**. Two sources on Syria provide most of the data – SHAM and SOHR – both of them include information that is unique. The analysis highlights that the most feasible strategy has been to take one of the sources as the baseline and engage in targeted strategies to supplement the data. Our analysis highlights that SOHR is the most viable source to serve as a baseline. This is a controversial choice: SOHR is not transparent enough 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.[23]

2. **Targeted enrichment to complement the baseline**. The analysis highlights that Liveuamap and SHAM provide unique information not contained in SOHR’s data. Liveuamap reports on unique event-types and actors. Sham is better regarding unique locations. While Liveuamap can be captured within reasonable amounts of time, the amount of data generated by Sham is so large that we only include areas

[23] Using SOHR means that most events will have an initial event reliability of 3.

where Sham is most different from SOHR. These areas are: Hama, Hasakeh and Rural Damascus.[24] This pragmatic choice comes 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, now or in the future, will allow for further improvement of the database.

3. **Targeted enrichment for known deficiencies in the data**. The analysis showed three common deficiencies. One, the underreporting of violence against civilians. Second, the underreporting of events in Al-Hasakeh and Raqqa and finally, the unclear identity of actors involved in remote violence (IED and airstrikes).[25] Based on this outcome, we sought additional public sources and approached specific organizations who are known to report on these deficiencies. After testing various organizations and Twitter feeds we ended up with an additional set of sources.

Table 9: Targeted strategy for Syria

Deficiency		Source	
-	Baseline	SOHR	
0	Targeted increase	Liveuamap	
		Sham (Hama, Hasakeh, Rural)	
1	VAC	SNHR	
		Undisclosed source 1	
2	Coverage of Al-Hasakeh & Raqqa	RBSS	
		Hawar News Agency[26]	
3	Unidentified actors	Airwars (Coalition airstrikes) [27]	
		UNSC – SG reports (Airstrikes)	

Conclusion and next steps

This report provides descriptive statistics on a pilot-project on the coverage of the Syrian conflict in three separate weeks. The purpose of this pilot was to devise a strategy to combine different data sources into one compressive, high quality database. The design uses the ACLED infrastructure and goes through a purposely designed two-step process to enrich data.

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

[25] Almasdar News (Pro-government) may be a good source to balance the somewhat opposition leaning nature of the data in the database. Addressing underreporting 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.

[26] 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).

[27] A test of inclusion of all ISW reports (detailing Russian/Israeli airstrikes), resulted in too few events.

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 organization critical of Assad regime.	Local sources through own reporting network of citizen journalist.	2012-present	www
Syria Deeply	New story-telling of conflict situations. Focus on political analysis, truce, 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, SNHR, Syria 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 of individuals who are dead, missing or 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 Islamic State and Syrian governments in Raqqa and surroundings.	Formerly most activists in Raqqa, recent reports draw on contacts outside Raqqa.	2014-present	English	www
Syrian Tracker	Records violations such as killing, 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
Committees for the Defense of Democracy,	Records individual deaths and casualties, including locations and names.	Claims use of social media but unclear exactly what types of sources and through	2010-present	Arabic	www

Freedoms and Human Rights in Syria		which procedure.			
Damascus Center for Human Rights Studies	Records individual casualties, and victims of extrajudicial killings, massacres, arbitrary detention, enforces 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 & 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 Human Rights Observatory (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
Syrian Network for Human Rights (SNHR)	Records detainees, deaths and attacks on vital facilities 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 organizations, Syrian research organizations, field hospitals, Local Councils/Coordination Committees, Local network (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 Centre	*Type*	Records 'conflict events' giving dates and exact locations. Events include bombings/shelling, clashes, IEDs/suicide bombings and territorial take-over	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.	Primary sources: ISW Syria team (local reporting	2013-present	www

		Occasionally other information such as troop movement or detailed city control maps.	network). Secondary sources: wikimapia, SHOR, Syria Direct, Sham News Network. Sources changed over time. Grading of source reliability.		
	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 source, 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.	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-crow 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 (theft, threat but also assault and attacks).	Information reported to INSO by its NGO members	2014-present	www
	Actors	General: Assad Regime, Local Ethnic Militias, Coalition, OAG (organized armed groups)			
IHS Conflict Monitor (Jane's)	*Type*	Territorial control, number 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

		SHAM	SOHR	SNHR
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			

Table 3: Governorate coverage

	SHAM	SOHR	SNHR	Lexisnexis	Airwars	ISW	SANA	Liveuamap
Riots/Protests	6	2						
Violence against civilians	16	34	110	5		2		4
Remote violence	898	822	183	55	33	27	59	124
Battles	281	478	146	87		28	24	64
Other	18	11		1		9		

Table 5: Event-type coverage

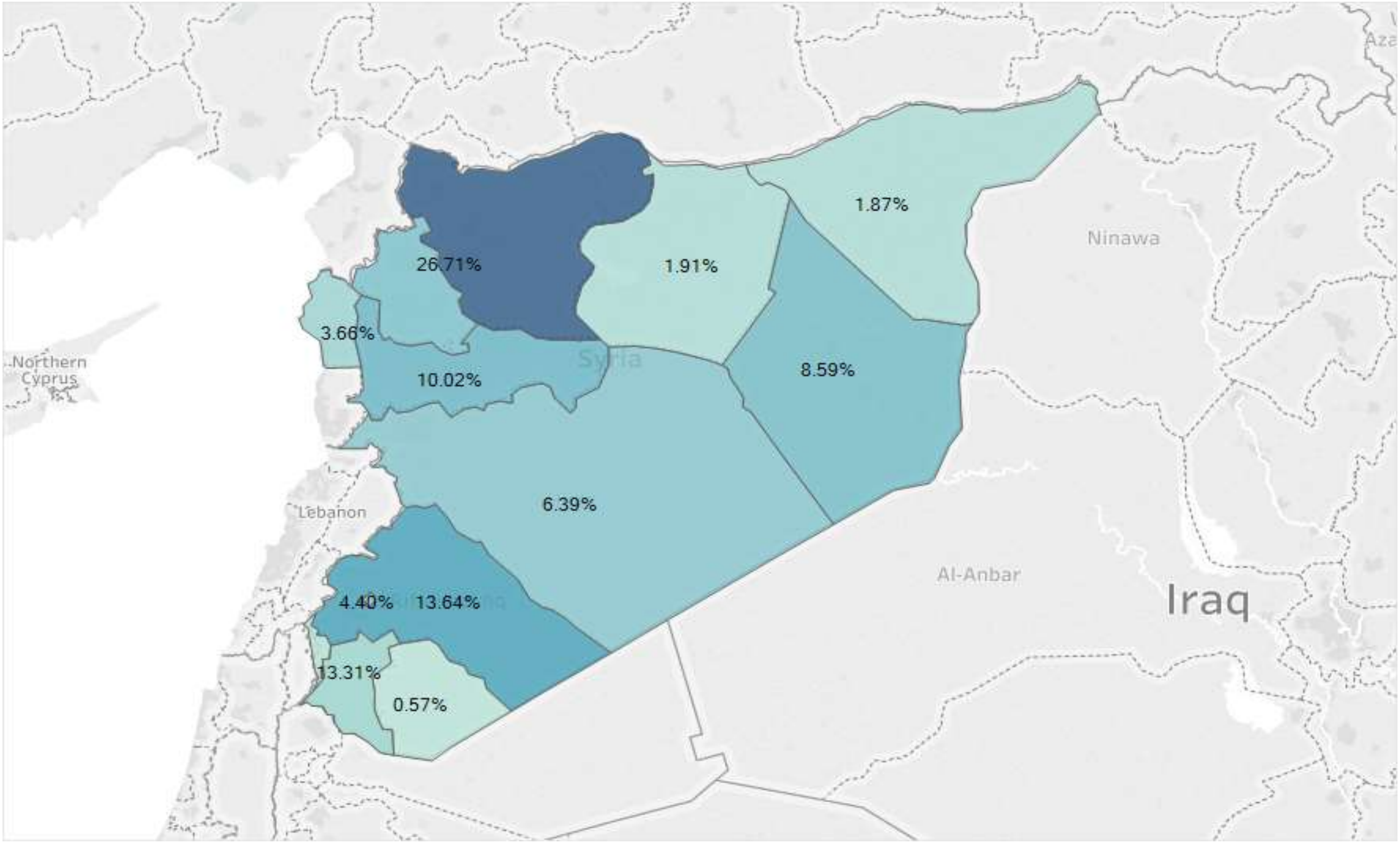


Figure 3: Coverage governorate

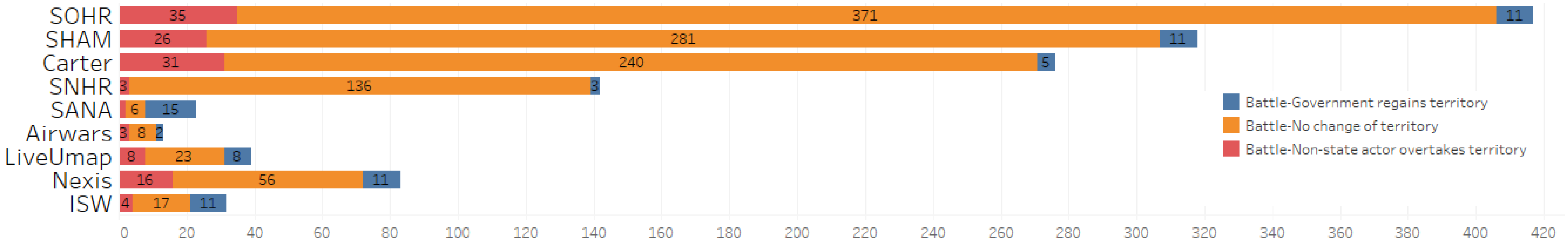


Figure 4: Territorial take-over

Annex 3: Reliability Sheet Syria

Source used: Carter Centre, Airwars, Liveuamap, Syrian Network for Human Rights (SNHR), Syrian Human Rights Observatory (SOHR), Undisclosed source 1, Sham News, RBSS, Hawar News Agency, UNSC – SG reports. **Sources not-used**: Sana, ISW, Lexisnexis, Syria Direct, Syrian Archive

Amount of violence captured: unclear (likely high).

Likely bias:	**Addressed**
Description of events follows the interest of organizations.	Yes
Organizations depict actors differently	Yes
Underreporting from rural areas	No
Underreporting with no road network	No
Underreporting when there is a great distance from the capital	Yes
Underreporting when there is no cell-phone coverage	No
Underreporting of low-impact and minor events	Partly
Underreporting where the organization has no network	Partly
Violence is better reported when there is a large population	Partly
The killing of 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 standardization; 2) storing / Open and transparent public methodology
SNHR	Validated	1) collection by recorders; 2) collect visual proof (2 sources); 3) database / crosschecking; 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	Credible	Journalist principles / No public methodology available
SOHR	Credible	Usage claimed 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 1	One-Source	Internal reporting / No public methodology available