Converting and Enriching Geo-annotated Event Data: Integrating Information for Ukraine Resilience

Manar Attar
Shuai Wang
Ronald Siebes
m.m.attar@student.vu.nl
{shuai.wang|r.m.siebes}@vu.nl
Vrije Universiteit Amsterdam
Amsterdam, the Netherlands

Eirik Kultorp
Amstelveen, the Netherlands
ekultorp@gmail.com

ABSTRACT

The mission of resilience of Ukrainian cities calls for international collaboration with the scientific community to increase the quality of information by identifying and integrating information from various news and social media sources. Linked Data technology can be used to unify, enrich, and integrate data from multiple sources. In our work, we focus on datasets about damaging events in Ukraine due to Russia's invasion since February 2022. We convert two selected datasets to Linked Data and enrich them with additional geospatial information. Following that, we present an algorithm for the detection of identical events from different datasets. Our pipeline makes it easy to convert and enrich datasets to integrated Linked Data. The resulting dataset consists of 10K reported events covering damage to hospitals, schools, roads, residential buildings, etc. Finally, we demonstrate in use cases how our dataset can be applied to different scenarios for resilience purposes.

CCS CONCEPTS

• **Information systems → Extraction, transformation and loading**; **Mediators and data integration**; **Geographic information systems**.

KEYWORDS

Linked open data, data integration, linked geospatial data, Ukraine resilience

ACM Reference Format:
Manar Attar, Shuai Wang, Ronald Siebes, and Eirik Kultorp. 2023. Converting and Enriching Geo-annotated Event Data: Integrating Information for Ukraine Resilience. In *The 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '23), November 13–16, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589132.3625580

SIGSPATIAL '23, November 13–16, 2023, Hamburg, Germany

ACM ISBN 979-8-4007-0168-9/23/11.
https://doi.org/10.1145/3589132.3625580

1 INTRODUCTION

The Russian-Ukrainian conflict since February 2022 damaged civilian infrastructure, facilities, and buildings, sparking a large displacement crisis in Europe. According to the Ministry of Health in Ukraine, 117 medical institutions have been destroyed [1]. In Kharkiv alone, the war has resulted in large-scale destruction of infrastructure with an estimation of more than 1,000 buildings destroyed, among which 700 are multi-storey apartment buildings but no longer habitable [1]. Rebuilding destroyed public facilities and social infrastructure is critical for the resilience of Ukraine. It is no easy task and will require international cooperation and coordination for reconstruction, including integration and management of datasets and resources of various kinds.

There are many open datasets that report the progress of the Russian-Ukrainian conflict from various perspectives. These datasets enable researchers and analysts to gain insights into the complex situation, ultimately contributing to the development of effective strategies to protect civilians, promote peace, estimate resources for projects for resilience, etc. WikiEvents [3] consists of entries automatically curated based on Wikipedia's Current Events portal.[1] Its NLP downstream pipeline extracts 21,275 events including around a thousand events about the Russian invasion of Ukraine. ACLED is a much larger dataset[2] with over one million events, including around 40,000 political violence events across Ukraine [4]. However, this dataset is mostly dedicated to military use with three-quarters of its events about shelling, artillery, and missile strikes. The Centre for Information Resilience (CIR) launched the Eyes on Russia (EoR)[3] project in January 2022 with the aim of gathering and verifying media related to Russia's invasion of Ukraine. The project's primary objective is to provide access to verified information through a database and an interactive map, benefiting journalists, non-governmental organizations (NGOs), policymakers, and the public. The interactive map displays relevant information such as the data source, a description of the event, location coordinates, and the extent of damage caused. Furthermore, it includes a variety of classes, including the country name, province, city, coordinates, date, damage level, and source of information. The Civilian Harm in Ukraine TimeMap (CH)[4] is a similar project that provides a comprehensive record of such incidents by including

[1] https://en.wikipedia.org/wiki/Portal:Current_events
[2] https://acleddata.com/2023/03/01/war-in-ukraine-one-year-on-nowhere-safe/, visited on 13th June, 2023.
[3] https://eyesonrussia.org/
[4] https://ukraine.bellingcat.com/

source links, precise location data determined by the Global Authentication Project and Bellingcat researchers, and a brief description based on visual evidence. Its structured data can be used for further analysis and research to understand better the impact of the conflict on civilians in Ukraine. EoR and CH are the two datasets selected for this study given their similar approach in generating and representing events. Both projects focus on damage reporting and serve as important resources for those seeking accurate and verified information to aid decisions for the resilience of Ukraine.

Accurate and complete documentation of the damage could benefit projects for resilience in multiple ways. Linked Data is structured data that can be interlinked with other data, which enables additional functions through semantic queries. While it is not common to use Linked Data in projects for resilience, past projects have demonstrated the use of Linked Data in decision-making in government, NGOs, and societal organizations. For example, the Brazilian government used ontologies and enriched data from various governments, resulting in a DBpedia-like Government Open Linked Data - DBGOldBr [6]. Our project explores how the transformation of event data into Linked Data facilitates an integrated and more complete description of events. For example, we consider the following event[5] in CH that happened on 7th March 2022. It was reported to have "Hospital destroyed by explosion". Its location information is "Izum, Kharkiv region" and it lacks information about the postal code.

While the above-mentioned datasets were initially designed for their platforms, one can take advantage of ontologies and Linked Data technologies to provide a unique representation of entities such as cities and provinces to reduce ambiguity, which could make it easier for integration and verification, and enable interoperability with datasets in other disciplines (e.g. economic and social-/historical- data). In this paper, we attempt to convert and unify structured geo-annotated datasets. More specifically, we convert two existing geo-annotated datasets dedicated to damage reporting in Ukraine to their corresponding representation as Linked Data. We propose a pipeline for the integration of datasets and demonstrate the use of the resulting dataset. Finally, we evaluate the quality of the integrated data and demonstrate its use by developing a web application that shows detailed information regarding damaged locations.

Our research question is: How to unify geo-annotated events in multiple datasets about damaging events? We answer this question by studying the following sub-research questions:

- **SRQ1:** How can we provide a unified representation of information in the datasets as Linked Data?
- **SRQ2:** How can we enrich the converted Linked Data with geospatial information?
- **SRQ3:** How can we integrate datasets by identifying and merging entities that describe the same events?
- **SRQ4:** What is the quality of the resulting unified data?

The research output of this paper includes 1) the converted datasets together with related resources; 2) an integrated dataset; 3) a pipeline with open source code that can be adapted to future datasets; 4) use cases with SPARQL queries.

[5]The event was extracted manually based on the Twitter post https://twitter.com/KyivIndependent/status/1501218105342763020.

This paper is organized as follows. Section 2 includes details of data pre-processing, unification, conversion, and enrichment. Section 3 outlines the design of the algorithm that detects duplicates in the selected datasets as part of the automated integration pipeline. We publish our dataset with evaluation in Section 4. Some use cases are included in Section 5. Finally, we discuss the limitations of our approach and present plans for future work in Section 6.

2 DATA PROCESSING

For our study, we selected two datasets: Eyes on Russia (EoR) and Civilian Harm (CH).[6] There are 9,308 and 1,105 events in EoR and CH, respectively. Both datasets have coordinates associated with every event. Entries in the datasets have been reviewed by volunteers and data curators. Some events have missing information. Despite that the datasets are bilingual, events are mostly in English with a few in Ukrainian. In the following subsections, we provide details of data conversion, enrichment, and unification.

2.1 Data Conversion

Our examination of the datasets shows that the fields and formats of reported events can vary significantly. This is partially due to the lack of use of controlled vocabularies and ontologies. Take the location information of CH for example, the event in Section 1 has location information "Izum, Kharkiv region". However, we observed other formats such as "Kharkiv", "Merefa, Kharkiv", as well as poorly formatted strings such as "\r\nZhytomyr", and mistakes such as "Kyiv region, Donetsk". To answer SRQ1, we select entities and relations from popular ontologies such as Schema.org[7], the Dublin Core[8], Simple Event Ontology[9], and the GeoNames[10] for a unique representation of (geo-)information of events. In addition, we also introduce some relations in our own namespace. Moreover, some specific information is not generic between datasets, e.g. violence level and the type of damage 'Civilian Infrastructure Damage'. We include such information in the comment (as the object of `rdfs:comment`) to be studied in future work.

We assign a Uniform Resource Identifier (URI) to each event. We model that each event is of type *Event* as in the Simple Event Ontology [5]. We noticed that many events were reported with an accurate date but not the exact time. In fact, many happened at exactly 00:00:00, which could be the default time setting. Therefore, we ignore the exact time of the event and take the day without the time. Following that, we use its coordinates and find its unique representation of province, city, and postal code in GeoNames. As for the example in Section 1, the reported province/region is Kharkiv. We retrieve Kharkiv's corresponding URI in GeoNames: http://sws.geonames.org/706483/. However, its postal code is still missing. This leads to the step of data enrichment in the next section.

2.2 Data Enrichment

It was noticed that some information is not explicitly provided but can be inferred. For example, the postal code can be retrieved by calling GeoNames' APIs. Recall our example in Section 1, the

[6]Both datasets were retrieved on 30th April 2023 from their official websites.
[7]https://schema.org/docs/schemas.html
[8]https://www.dublincore.org/
[9]https://semanticweb.cs.vu.nl/2009/11/sem/
[10]https://www.geonames.org/

	EoR			CH		
	O	**CE**	**comment**	**O**	**CE**	**comment**
country	9308	9308	obtained GeoNames' country URI using the string	0	1105	obtained GeoNames' country URI using the coordinates
city	9308	9308	obtained GeoNames' city URI using the string and coordinates	unknown	1105	converted from string to GeoNames' city URI or retrieved using coordinates
province	9308	9308	25 were manually corrected due to incorrect spelling	unknown	1105	for inconsistent representation, their province was obtained as GeoNames' province URI by using their coordinates
date	9308	9308	converted from string to date:xsd format	1105	1105	converted from string to date:xsd
coordinates	9308	9308	added as GeoCoordinates format	1105	1105	added as GeoCoordinates format
postal code	0	9223	retrieved from GeoNames using the coordinates (85 entries do not have a corresponding postal code in GeoNames)	0	1105	retrieved from GeoNames using the coordinates
description	9306	9306	two events lack description.	1105	1105	kept original
URL	9308	9308		1057	1057	
violence level	9296	0	the violence level was left as comments due to lack of standards and definition	0	0	CH does not have the value violence level
#events	9308	9308		1105	1105	

Table 1: Comparison of the Eyes on Russia and Civilian Harm datasets entries (O: The original dataset before and after processing, CE: the dataset after conversion and enrichment.)

missing information postal code is 64305. Not all information was represented correctly. Take EoR for example, only 8,884 events have their city information formatted correctly and found in GeoNames. Another 368 associated strings were about villages, towns, local neighborhoods, or other names that do not exist as cities using GeoNames. 56 events have none of the corresponding information mentioned above. Therefore, we retrieved this information from their coordinates in GeoNames. Difficulty due to spelling errors and multilingual cases were manually resolved. Table 1 presents a summary of conversion and enrichment.

3 DATA INTEGRATION

As for SRQ3, our manual examination shows that cases where one event was reported two or multiple times are very rare. Thus, we rely on the Unique Name Assumption for both datasets: no event was reported twice at a close distance in the same dataset. Algorithm 1 takes into consideration the distance of events from two datasets and their description. We manually fine-tuned all the parameters by experimenting with results that gave reasonable outputs[11]. The output of the following algorithm consists of 1) pairs of events that we consider potentially identical (denoted S) and 2) pairs of events that are close to each other but not identical (denoted T).[12] As for string similarity, we took advantage of the *SequenceMatcher* function in the *difflib* Python package.[13] Other sequence comparison methods will be explored in the future.

Our manual examination shows that the coordinates of reported identical events about an 'area' could be some distance apart. Therefore, we take two different strategies for areas and other cases separately. We consider a broader radius of 2km for events about

for *each pair of events* (i, j) *with identical city and date* **do**
  $d \leftarrow$ the distance between i and j;
  $s \leftarrow$ the similarity between the description of i and j
  if *i and j are backed by the same social media link and* $s > 0.55$ *and* $d < 2(km)$ **then**
    add (i, j) to S
  else
    if *'area' is in the description of i or j* **then**
      if $s > 0.75$ *and* $d < 2(km)$ **then**
        add (i, j) to S
      else
        add (i, j) to T
      end
    end
    if *keywords such as 'school', 'hospital' are in the description of i or j* **then**
      if $s > 0.55$ *and* $d < 1(km)$ **then**
        add (i, j) to S
      else
        add (i, j) to T
      end
    end
  end
end

Algorithm 1: Data integration using distance, description, and associated link to social media content

'area'. For other cases, we consider only the keywords about theater, church, school, hospital, building, house, flat, station, etc. Other reported events such as military operations are not considered.

We identified 206 pairs of events and we associate each with a new event URI that represents their integration. Moreover, we

[11] Details of the selection of parameters are included in the supplementary material.
[12] Other pairs of events are stored for manual examination in future work.
[13] https://docs.python.org/3/library/difflib.html

introduce an additional `hasPrimarySource` relation in our namespace for the primary source (the event with richer information). Overall, we included 10,207 events in the integrated dataset.

4 EVALUATION AND PUBLICATION

Finally, for SR4, we assess the quality of our algorithm and the resulting datasets. For the former, we created a questionnaire that consists of randomly selected 10 pairs of events from S (pairs of events considered identical) with 10 additional pairs of events selected from T (pairs of events from the same city, on the same day, and close to each other but not considered identical). We received 6 valid submissions by the deadline.[14] By assigning a number to each answer (2 for 'Very likely', 1 for 'Likely', 0 for 'Unsure', -1 for 'Unlikely', and -2 for 'Very unlikely'), our analysis of the results indicates that pairs of events considered identical by our algorithm have an average of 1.38 (between 'Likely' and 'Very likely' to be identical). In comparison, that of other events is -0.45 (between 'Unsure' and 'Unlikely'). This shows that our algorithm has good precision while those we decide to leave out remain unsure.

Our datasets are hosted on the TriplyDB platform[15], an RDF datastore with various data visualizations. Passing it through its data processing pipeline ensures syntactic correctness. As sanity checks of our data, we run SPARQL queries to manually validate the ranges of data points on temporal and spatial dimensions. Some missing entries and multilingual cases were manually handled. The corresponding converted datasets, the SPARQL queries, a demo video, the code, the questionnaire, as well as other supplementary material used are available on GitHub[16].

5 USE CASES

Use Case 1: Events visualization. As a demonstration of the use of our integrated dataset, Figure 1 presents the result of a SPARQL query that retrieves events in Kherson in the integrated datasets between October 1^{st} 2022 and February 28^{th}, 2023.

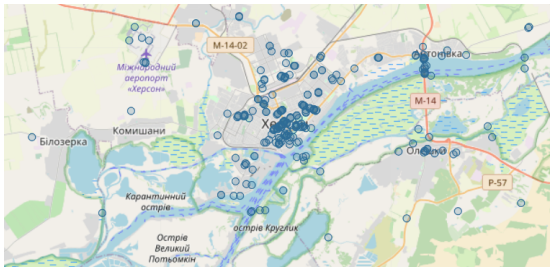

Figure 1: Events in Kherson

Use Case 2: Timelapse of damaging events about schools and hospitals. Figure 2 illustrates dates and their corresponding number of events about schools, universities, and hospitals between August 1^{st}, 2022 and April 30^{th}, 2023. This information could be used for the estimation of budgets for rebuild/repair.

[14]The questionnaire and results are included in the supplementary material.
[15]https://triplydb.com/linked4resilience/.
[16]https://github.com/LinkedData4Resilience/linked-data. Given that the resulting integrated dataset could be used for unintended purposes, it is accessible upon request only. Further updates and more use cases are on the website: https://linked4resilience.eu/.

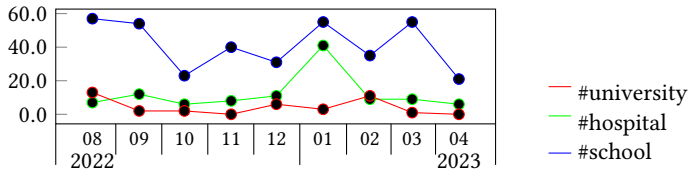

Figure 2: Timelapse of events about public facilities

6 DISCUSSION AND CONCLUSION

This paper presents how existing datasets about damage reporting in Ukraine can be converted to Linked Data. An algorithm was designed for the automatic detection of identical events, which was used to integrate events from two datasets. Our approach reduces ambiguity and enables the enrichment of events with information from other linked open data sources. Finally, we demonstrate how the resulting integrated dataset can be used for resilience purposes.

Some invalid links to social media content were detected, including broken/missing links and links to content that requires access permission. Our examination shows that 1.1% and 11.6% are invalid in EOR and CH, respectively, which indicates that the information gathered from social media platforms may not be reliable or complete. This problem is more present for CH. Similar issues have been discussed in some previous research [2]. Further assessment and validation are required if better accuracy is essential.

The resulting dataset could be further enriched with information about the type of buildings, schools, etc. Moreover, the labels such as cities, and provinces could be enriched with multilingual information. Our approach can be further extended to include additional datasets, such as ACLED [4] and WikiEvents [3] to construct a more inclusive and accurate estimation of resilience needs. Our approach provides insights into integrating multiple sources about damage, cultural heritage, shelters, traffic, and other related information.

Given the small size of datasets, although our approach is designed for static data, it has the potential to be used for continuous data streams and can be adapted for other types of datasets of geo-annotated events or the resilience of other countries.

Acknowledgement. The authors thank Tianyang Lu, Zhisheng Huang, Igor Potapov, Olexandr Konovalov, and volunteers for their help.

REFERENCES

[1] Dmytro Chumachenko and Tetyana Chumachenko. 2022. Ukraine war: The humanitarian crisis in Kharkiv. *BMJ* 376 (2022). https://doi.org/10.1136/bmj.o796
[2] Kate Crawford and Megan Finn. 2015. The limits of crisis data: analytical and ethical challenges of using social and mobile data to understand disasters. *GeoJournal* 80 (2015), 491–502.
[3] Vasilis Kopsachilis, Nikos Vachtsavanis, and Michail Vaitis. 2023. WikiEvents - A Novel Resource for NLP Downstream Tasks. In *Proceedings of the 5th International Workshop on Semantic Methods for Events and Stories, SEMMES, 2023, Hersonissos, Greece, May 29th, 2023 (CEUR Workshop Proceedings).*
[4] Clionadh Raleigh, rew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing ACLED: An armed conflict location and event dataset. *Journal of peace research* 47, 5 (2010), 651–660.
[5] Willem Robert Van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. 2011. Design and use of the Simple Event Model (SEM). *Journal of Web Semantics* 9, 2 (2011), 128–136.
[6] Marcio Victorino, Maristela Terto de Holanda, Edison Ishikawa, Edgard Costa Oliveira, and Sammohan Chhetri. 2018. Transforming Open Data to Linked Open Data Using Ontologies for Information Organization in Big Data Environments of the Brazilian Government: the Brazilian Database Government Open Linked Data–DBgoldbr. *Knowledge Organization: KO* 45, 6 (2018), 443.