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

BSc Thesis *(Afstudeerscriptie)*

written by

Manar Attar
(born March 27th, 1997 in Damascus, Syria)

under the supervision of **Mr. Shuai Wang** and **Dr. Ronald Siebes**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

Bachelor in Artificial Intelligence
Intelligent Systems Track

at the *Vrije Universiteit Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
July 13th, 2023	Prof. Jacco van Ossenbruggen (second reader)

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.

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

Contents

Chapter 1

Introduction

Linked Data is structured data that can be interlinked with other data, which enables additional functions through semantic queries. While the use of linked data in resilience projects is not yet common, previous projects have demonstrated its utility in decision-making within governmental bodies, NGOs, and societal organizations. Notably, the Brazilian government employed ontologies and enriched data from various sources, resulting in a DBpedia-like Government Open Linked Data - DB-GOldBr [16].

In this thesis, we purpose to explore how the transformation of event data into Linked Data facilitates an integrated and more comprehensive description of war events in Ukraine. Providing an extensive record of the destruction caused by the Russian-Ukrainian conflict. The extent of the damage, including the destruction of 117 medical institutions reported by the Ministry of Health[2] and over 1,000 buildings underscores the urgency and scale of the reconstruction process. Promoting resilience in Ukraine requires not only the physical reconstruction of public facilities and social infrastructure but also the integration and management of diverse datasets and resources.

To illustrate, we consider an event[1] from the CH dataset that occurred on March 7th, 2022, reported as "Hospital destroyed by explosion" in the location of "Izum, Kharkiv region." However, this event lacks information about the postal code. By leveraging ontologies and linked data technologies, we seek to reduce ambiguity, enable easier integration, and foster interoperability with datasets from other disciplines, such as economic and social/historical data.

To address these needs, this thesis converts and unifies structured geo-annotated datasets, specifically two existing datasets focused on damage reporting in Ukraine, into their corresponding format in linked data. We propose a comprehensive pipeline for dataset integration and demonstrate the practical use of the resulting dataset. Additionally, we evaluate the quality of the integrated data and showcase its application by developing use cases and a demo that provides detailed information about damaged locations.

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

Our research question is: How to unify geo-annotated events in multiple datasets? 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 do 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 the thesis is as follows:

1. the converted dataset 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.

This thesis is organized as follows. Chapter 2 investigates related approaches to the representation of damaging and resilience information. Chapter 3 compares datasets from various resource and select the datasets to be studied in this thesis based on criteria. Section 3.1 explains the conversion process then in Sections 3.2 and 3.3 includes details of data enrichment and multilingual representation. Following that, we convert the selected datasets to linked data and integrate them in Chapter 4. Further, we indicate the evaluation process in chapter 5. To demonstrate the use of our integrated dataset, we designed some use cases and demos. They are included in Chapter 6. Finally, we discuss the limitations of our approach, alternative approaches, as well as future work in Chapter 7.

Chapter 2

Related Work

There are various open datasets that provide valuable insights into the Russo-Ukrainian conflict from different perspectives. Two prominent datasets in this domain are WikiEvents [6] and ACLED [10]. WikiEvents is a collection 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 related to the Russian invasion of Ukraine. The dataset enables researchers and analysts to gain insights into the complex situation, contributing to the development of effective strategies to protect civilians, promote peace, and estimate resources for resilience projects.

ACLED, on the other hand, is a much larger dataset[2] with over one million events, including approximately 40,000 political violence events across Ukraine. While ACLED provides a vast amount of data, it primarily focuses on military aspects, with a significant portion of its events centered around shelling, artillery, and missile strikes.

In response to the conflict, the Centre for Information Resilience (CIR)[3] launched the Eyes on Russia (EoR)[3] project in January 2022. EoR aims to gather and verify media content related to Russia's invasion of Ukraine, providing access to verified information through a database and an interactive map. This project is valuable for journalists, NGOs, policymakers, and the public, offering insights into the conflict and its impact.

Another noteworthy project is the Civilian Harm in Ukraine TimeMap(CH)[4], which provides a descriptive record of incidents, including source links, precise location data, and descriptions based on visual evidence. The structured data provided by CH facilitates a better understanding of the impact of the conflict on Ukrainian civilians.

While these datasets were initially designed for their respective platforms,

[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/

our project targets to leverage ontologies and linked data technologies to provide a unified representation of entities such as cities and provinces, reducing ambiguity and enabling easier integration and interoperability with datasets in other disciplines [5].

By reviewing the existing literature and datasets related to damage reporting and resilience in Ukraine, we can identify gaps and opportunities for our research, which focuses on the conversion, unification, and enrichment of geo-annotated datasets to facilitate integrated and more complete event descriptions.

In addition to the aforementioned datasets, several studies have explored the use of linked open data and ethical considerations in crisis events and big data environments.

One relevant article focuses on transforming open data into linked open data using ontologies in the context of the Brazilian government's DBpedia-like Government Open Linked Data (DBGOldBr)[16]. initiative. The authors highlight the challenges faced by the Brazilian government in organizing and providing access to vast amounts of data. They explain how the use of linked open data and ontologies can help overcome these challenges by improving data integration, interoperability, and reuse. The article describes the ontology development process and provides examples of how DBGOldBr was utilized to link and enrich data from different government agencies.

Another study addresses the limitations of using social media and phone data to understand people's responses during crises [4]. Using examples of Hurricane Sandy and the Haiti Earthquake, the authors demonstrate how relying solely on platforms like Twitter and crowdsourced text messages raises questions about the representativeness of the data. The article emphasizes the importance of being aware of ethical frameworks and understanding the limitations of the data to obtain a more accurate understanding of crisis events. Furthermore, it highlights the need for researchers to consider the impact of their work on privacy and consent.

Ethical considerations in using machine learning models on social media data are explored in a separate paper [7]. The report provides an overview of social media's evolution and discusses the various ethical risks associated with employing machine learning techniques for data collection. Privacy concerns, bias in algorithms, and the manipulation of information are among the ethical challenges discussed. The study emphasizes the responsible collection, use, and sharing of data, as well as the importance of transparent and fair decision-making processes.

These studies demonstrate the importance of incorporating linked open data approaches, considering ethical frameworks, and understanding the limitations of data sources in crisis events and big data environments. By leveraging linked data and addressing ethical concerns, researchers can enhance the integration, analysis, and interpretation of data in the context of resilience projects, leading to more informed decision-making processes.

Chapter 3

Data Processing

EyesOnRussia: The EyesOnRussia project has generated a dataset that includes a variety of classes, including the country name, province, city, coordinates, date, damage level, and source of information. This dataset is provided in JSON format, and it is intended to support further analysis and research on the conflict between Russia and Ukraine. Overall, the Eyes on Russia project serves as an important resource for those seeking accurate and verified information about the ongoing conflict, and its availability in a structured dataset further enables researchers to gain insights into the conflict's complexities. [1]

Civilian Harm: The Civilian Harm dataset contains several classes, including city, coordinates, date, and source of information, which are available in JSON or CSV format. This structured data can be used for further analysis and research to understand better the impact of the conflict on civilians in Ukraine. Overall, the Civilian Harm in Ukraine project plays an essential role in documenting and raising awareness of the potential harm to civilians in the ongoing conflict in the country. The project's structured data further enables researchers and analysts to gain insights into the complexities of the situation, ultimately contributing to the development of effective strategies to protect civilians and promote peace. [2]

leedrake5-Russia-Ukraine[9]: The Russia-Ukraine War tracker is a tool that leverages data from Oryx's site to visualize equipment losses since Russia's invasion of Ukraine on February 24th. This tracker provides independent verification of the destroyed vehicles and equipment for which photo or videographic evidence is available. The data is sourced from a public Google sheet that is updated based on the latest information available for each day. The classes of the Russia-Ukraine War tracker include unit loss by country and date, which are presented in the Google Sheets format. This structured data allows for the analysis of the losses incurred by various military units over time and provides insights into the ongoing conflict between Russia and Ukraine. Overall, the Russia-Ukraine War tracker serves as a useful resource for those seeking to un-

[1] Eyes on Russia data set was retrieved last on 30th of April 2023

[2] Civilian Harm data set was retrieved last on 30th of April 2023

derstand the impact of the conflict on military equipment and operations. The availability of structured data in the Google Sheets format further facilitates the analysis of losses incurred by various military units and can help inform strategic decision-making.

StandWithUkraine[17]: StandWithUkraine is an open data service that provides free and easy access to information about Ukraine. The service offers a simple interface and an API for developers to access the data, making it a valuable resource for researchers, analysts, and other potential users interested in understanding various aspects of Ukraine. The Ukrainian Data Hub, which is part of StandWithUkraine, provides a repository of structured data on various topics, including demographics, economics, and social issues. This data can be used to identify trends, develop insights, and support evidence-based decision-making. Additionally, StandWithUkraine maintains a GitHub repository with statistics and information about humanitarian numbers, providing a valuable resource for those interested in the humanitarian situation in Ukraine. Overall, StandWithUkraine plays a critical role in promoting open data and transparency in Ukraine, making valuable information accessible to a wide range of users. The service's simple interface and free access via API and GitHub repository further facilitate the use of data to support research and analysis.

Russia block list[15]: The Russia block list is a compilation of companies that have suspended their operations in Russia due to various reasons. These reasons may include economic sanctions, political instability, or other factors that have made it difficult for these companies to continue operating in Russia. The block list provides a valuable resource for those interested in understanding the impact of economic and political factors on businesses operating in Russia. By compiling information about companies that have suspended their operations, the block list can help identify patterns and trends in the business environment in Russia and inform strategic decision-making. Overall, the Russia block list serves as an important tool for researchers, analysts, and other stakeholders interested in understanding the impact of various factors on businesses operating in Russia. By providing a list of companies that have suspended their operations, the block list can help shed light on the challenges and opportunities facing businesses in this complex and dynamic environment.

ACLED: Description: The ACLED project collects data on various forms of political violence, including armed conflict, riots, protests, and non-violent campaigns, among others. The data includes information on the location, date, actors involved, and fatalities and injuries resulting from the event. The project also collects data on the type of violence used, such as bombings, assassinations, and torture, as well as the targets of the violence, such as civilians, military personnel, or political leaders. Classes: Country name, province, city, coordinates, date, and the source of info in JSON format.

Decision to include or exclude: A Criteria for inclusion and exclusion has been defined to judge whether a dataset is useful to include in our project, our main focus is to describe war events happening in Ukraine as linked data, therefore the data set we will choose to include must have at least the following classes: A description of the event. The coordinates of the event. The source of the event The date of the event The availability of those classes makes it possible to convert the open data to open-linked data by the use of ontologies.

Include: Based on the defined criteria, the EyesOnRussia and Civilian Harm datasets contain a sufficient set of classes.

Additionally, we will retrieve address information such as province and postal code using coordinates via geocoding services. Geocoding is the process of converting a location's coordinates (latitude and longitude) into a human-readable address.

Exclude: Based on the defined criteria, Leedrake5-Russia-Ukraine and Stand with Ukraine data sets will be excluded, although they provide such interesting facts and information on the war in Ukraine, these datasets are outside of the scope of the project as they do not provide a description of the events happening nor a detailed location. These datasets provide statistical data about the war including humanitarian and refugee records that may be integrated in future work.

3.1 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 vocabulary and ontology. Take the location information of CH for example, the event in Chapter 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`[3], the Dublin Core[4], Simple Event Ontology[5], and the GeoNames[6] 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 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 [14]. 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 Chapter 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

[3] `https://schema.org/docs/schemas.html`

[4] `https://www.dublincore.org/`

[5] `https://semanticweb.cs.vu.nl/2009/11/sem/`

[6] `https://www.geonames.org/`

of data enrichment in the next section.

	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

Table 3.1: Comparison of old vs new Eyes on Russia and Civilian harm datasets entries (O: The original dataset, CE: the dataset after conversion and enrichment.)

3.2 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 Chapter 1, the missing information postal code is 64305. Not all information was represented correctly. Take EoR for example, only 8884 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.

3.3 Multilingual Information

In this thesis project using linked data, multilingual information is vital to accommodate collaboration across countries, improve usability for users, and enhance interoperability. Multilingual information enables diverse users to contribute effectively and understand project intricacies. By incorporating multiple languages, the project becomes more accessible, compatible and fosters a variety of resilience strategies that transcend language barriers. This inclusive approach promotes resilience and facilitates knowledge sharing among global communities[11]. Figure 3.1 displays multilingual labels for the city of Kupyansk, showcasing the diverse linguistic representation in our data

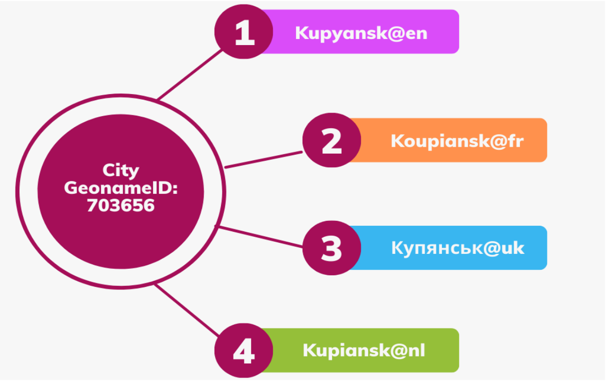


Figure 3.1: Multilingual representation of the Kupyansk City

To achieve the enrichment of data with multilingual labels, the Geonames web service API was employed. The process involved the following steps. Firstly, the geonameID obtained previously was utilized as a reference. Secondly, requests were made to the Geonames API, specifying the desired languages (such as French, English, Ukrainian, and Dutch) in the API calls. The API responded with the corresponding city names in the requested languages. These multilingual city names were then saved in a JSON file for future use during data conversion to RDF triples. Finally, the saved labels were incorporated into the data set, associating each city with its respective multilingual names. This enabled the data to be enriched with multilingual labels, providing language-specific information for further analysis.

Chapter 4

Data Integration

Next, we study SRQ3. Our manual examination shows that cases where one event was reported two or multiple times are very rare. Therefore, 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. 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).[1] As for string similarity, we took advantage of the *SequenceMatcher* function in the *difflib* Python package.[2] 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 'area'. For other cases, we consider only the keywords about theater, church, school, hospital, building, house, flat, station, etc. Since our dataset focuses mostly on damage reporting, other reported events such as military operations are not considered for merging and will be filtered out in the future.

We identified 206 pairs of events. For each pair, we associate them with a new event URI that represents the integration of these two events. Moreover, we introduce an additional `hasPrimarySource` relation in our namespace for the primary source of the event (the event with richer information). Recall that there are 9,308 and 1,105 events for EoR and CH, respectively. Overall, we include 10,207 events in the integrated dataset.

[1] Other pairs of events are stored for manual examination in future work.

[2] https://docs.python.org/3/library/difflib.html

```
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

Chapter 5

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.[1] By assigning a number to each answer[2], 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.[3] Passing it through TriplyDB's 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. The original datasets of CH and EoR, their corresponding converted datasets, SPARQL queries, and other resources used are available upon request.[4] Some missing entries and multilingual cases were manually handled. All the code is available on Github.[5]

[1] The questionnaire and results are included in the supplementary material.

[2] 2 for 'Very likely', 1 for 'Likely', 0 for 'Unsure', -1 for 'Unlikely', and -2 for 'Very unlikely'.

[3] https://triplydb.com/linked4resilience/. The integrated dataset is accessible upon request.

[4] https://figshare.com/s/6ca029692f0adf0f5deaA demo showing the use of the datasets.

[5] https://github.com/LinkedData4Resilience the data used in this project could be accessed on this GitHub reposotory

Chapter 6

Use Cases

In this section, we demonstrate the practical application of our data by designing SPARQL queries corresponding to some resilience aspects. Use case 6.1 highlights the value of visualizing events in Kherson on a map, and Use case 6.2 analyzes the multilingual representation of labels in the integrated dataset. Additionally, Use case 6.3 showcases the insights gained from timelapse visualizations of events about public facilities, and Use case 6.4 gives the analysis of monthly attacked regions. Furthermore, Use case 6.5 indicates the examination of the ratio of children's deaths to monthly attacks offers valuable insights into the humanitarian impact of the conflict. Together, these use cases illustrate the diverse applications of the integration of data to describe the various dimensions of the conflict in Ukraine.

6.1 Use Case 1: Events visualization

As a demonstration[1] of the use of our integrated dataset, Figure 6.1 showcases the outcome of a SPARQL query that retrieves events in Kherson within the integrated datasets from 1st October 2022 to 28th February 2023, which are then visualized on a map. By mapping these events, it becomes possible to identify patterns, concentrations, and trends on the map. The visual representation aids researchers and policymakers in gaining valuable insights into the spatial distribution of these events, thus facilitating better understanding and analysis.

This use case exemplifies the practical application of integrated datasets and the geographic visualization. The SPARQL query 6.1 aims to retrieve the location of the events and display them on the map using the primary sources predict, location, date, and geospatial coordinates within a specified timeframe. The query generates a WKT representation to support the visualization, and integration of the retrieved data.

[1] A demo showing the use of the datasets at `https://figshare.com/s/6ca029692f0adf0f5dea`.

```
PREFIX sdo: <https://schema.org/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX purl: <http://purl.org/dc/terms/>

SELECT ?wkt WHERE {
  ?aggregateEvent <https://linked4resilience.eu/ontology/
      hasPrimarySource> ?event .
  ?event sdo:location ?loc.
  ?event purl:date ?date .
  ?loc sdo:geo ?geo.
  ?geo sdo:latitude ?lat .
  ?geo sdo:longitude ?lng .
  filter(?startDate <= ?date)
  filter(?date <= ?endDate)
  bind(
    strdt(
      concat("POINT(",str(?lng),"␣",str(?lat),")") ,
      <http://www.opengis.net/ont/geosparql#wktLiteral>
    )
      as ?wkt)
}
```

Listing 6.1: SPARQL query for damaging events map

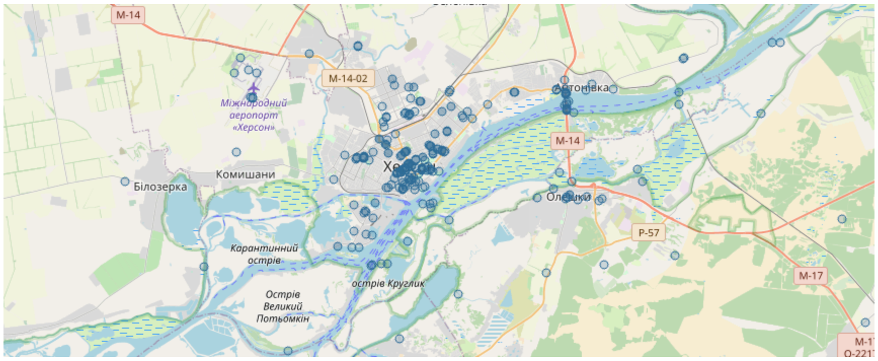


Figure 6.1: Events in Kherson

6.2 Use Case 2: Keyword damaging events

We developed a use case to showcase the timelapse of damaging events specifically related to schools and hospitals. Figure 6.2 presents a visual representation of the dates and the corresponding number of events concerning schools, universities, and hospitals during the period from 1st February 2022 to 30th April

2023. This timelapse provides an overview of the frequency and distribution of damaging events targeting educational and healthcare facilities over the specified timeframe. By observing the patterns and fluctuations depicted in the visualization, we can gain insights into the impact and severity of the conflict on these vital institutions.

The SPARQL query 6.2 aims to retrieve and calculate the maximum number of events that occurred in each month within a specified date range using a filter for a keyword. The query provides focused insights into event frequencies and patterns.

```
prefix dct: <http://purl.org/dc/terms/>

select ?monthyear (max(?numEvents) as ?numevts) where {
  {
    values ?monthyear {
      "2022-02" "2022-03" "2022-04" "2022-05" "2022-06" "
          2022-07"
      "2022-08" "2022-09" "2022-10" "2022-11" "2022-12" "
          2023-01"
      "2023-02" "2023-03" "2023-04"
    }
    bind(0 as ?numEvents)
  }union
  {
    select (count(distinct ?aggregateEvent) as ?numEvents) ?
        monthyear where {
      values ?monthyear {
        "2022-02" "2022-03" "2022-04" "2022-05" "2022-06" "
            2022-07"
        "2022-08" "2022-09" "2022-10" "2022-11" "2022-12" "
            2023-01"
        "2023-02" "2023-03" "2023-04"
      }
      ?aggregateEvent <https://linked4resilience.eu/ontology/
          hasPrimarySource> ?event .
      ?event ?predicate ?object .
      filter(isliteral(?object)) .
      filter(contains(lcase(str(?object)),lcase(?searchTerm)))
      ?event dct:date ?time .
      bind(replace(replace(replace(concat('0',str(month(?time)
          )),'010','10'),'011','11'),'012','12') as ?monthstr)
      bind (concat(year(?time), '-', ?monthstr) as ?
          computed_monthyear)
      filter(?computed_monthyear = ?monthyear)
    }
  }
} order by ?monthyear
```

Listing 6.2: SPARQL query for keyword damaged events time series

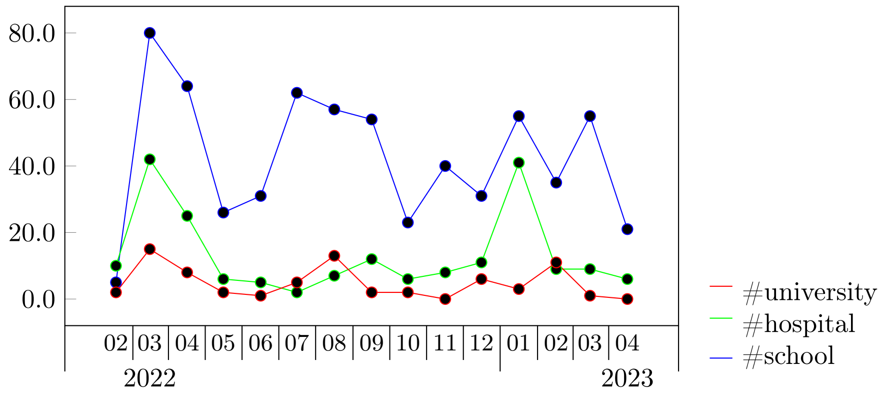


Figure 6.2: Timelapse of events about public facilities

6.3 Use Case 3: Multilingual representation of labels

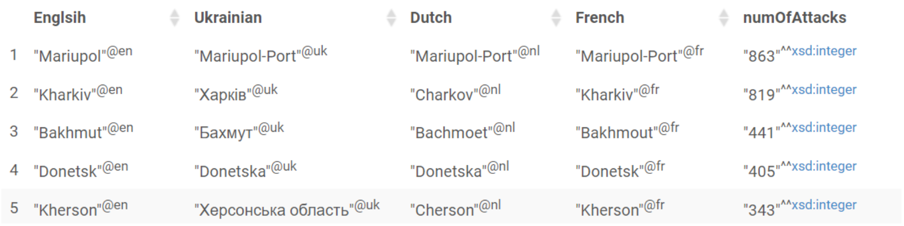
Incorporating multilingual information in a resilience project utilizing linked data is crucial for effective international collaboration, enhancing usability for many users, and improving interoperability. This holistic approach ensures that potential users of the data can actively participate, comprehend, and contribute to the project, ultimately fostering more resilient and inclusive communities. Figure 6.3 displays multilingual labels for the city of Kupyansk, showcasing the diverse linguistic representation in our data

The SPARQL query 6.3aims to retrieve distinct city name variations in English, Ukrainian, Dutch, and French. The query applies language filters, counts the occurrences of city name combinations, groups the results, and presents the top five combinations with the highest occurrence count.

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX l4r_o: <https://linked4resilience.eu/ontology/>

SELECT distinct ?Englsih ?Ukrainian ?Dutch ?French(count(?sub)
    as ?numOccurences) WHERE {
  ?sub l4r_o:cityName ?Englsih .
  ?sub l4r_o:cityName ?Ukrainian .
  ?sub l4r_o:cityName ?Dutch .
  ?sub l4r_o:cityName ?French .
  filter(LANG(?Englsih) = "en")
  filter(LANG(?Ukrainian) = "uk")
  filter(LANG(?Dutch) = "nl")
  filter(LANG(?French) = "fr")
} group by ?Englsih ?Ukrainian ?Dutch ?French order by desc(?
    numOccurences)
LIMIT 5
```

Listing 6.3: SPARQL query for representing cities in multilingual labels



	Englsih	Ukrainian	Dutch	French	numOfAttacks
1	"Mariupol"@en	"Mariupol-Port"@uk	"Mariupol-Port"@nl	"Mariupol-Port"@fr	"863"^^xsd:integer
2	"Kharkiv"@en	"Харків"@uk	"Charkov"@nl	"Kharkiv"@fr	"819"^^xsd:integer
3	"Bakhmut"@en	"Бахмут"@uk	"Bachmoet"@nl	"Bakhmout"@fr	"441"^^xsd:integer
4	"Donetsk"@en	"Donetska"@uk	"Donetska"@nl	"Donetsk"@fr	"405"^^xsd:integer
5	"Kherson"@en	"Херсонська область"@uk	"Cherson"@nl	"Kherson"@fr	"343"^^xsd:integer

Figure 6.3: YASGUI displaying the top 5 attacked cities in English, Ukrainian, Dutch, and French between February 2022 and April 2023

6.4 Use Case 4: Monthly attacked Regions

In this use case, we highlight the three most attacked regions in Ukraine each month. By creating a timeline shown in Figure 6.4 starting in February 2022 until December 2022, these results provide insights into resilience needs as the findings inform policymakers, researchers, and humanitarian organizations about areas requiring targeted support and intervention. This use case contributes to understanding resilience in Ukraine and guides further research and policy development in conflict resilience.

The SPARQL query 6.4 aims to retrieve distinct address regions. The query applies date filtering, counts the occurrences of each address region, groups the results, and presents the top three regions with the highest occurrence count. The analysis provides insights into the prevalence and distribution of address regions within the dataset, contributing to a better understanding of the geographic aspects of the events being studied.

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX l4r_o: <https://linked4resilience.eu/ontology/>
PREFIX purl: <http://purl.org/dc/terms/>

SELECT distinct ?obj (datatype(?obj) as ?dt)  (count(?sub) as
    ?numOccurences) WHERE {
  ?sub l4r_o:addressRegion ?obj .
  ?sub purl:date ?date .
  filter(?startDate <= ?date)
  filter(?date < ?endDate)
} group by ?obj  order by desc(?numOccurences)
LIMIT 3
```

Listing 6.4: SPARQL query for getting the most attacked region during certain period of time

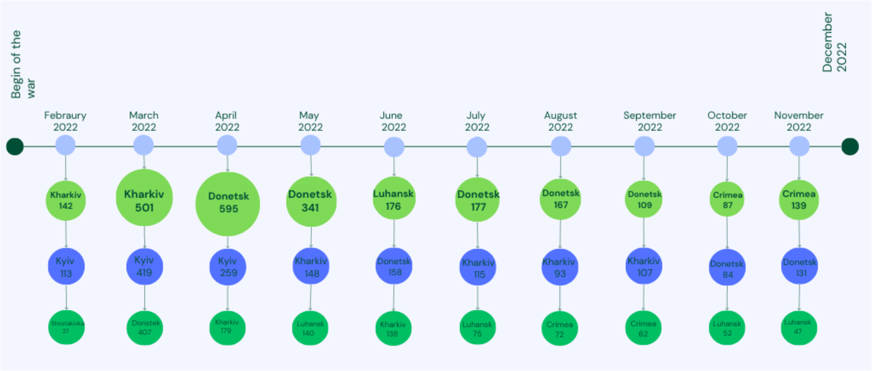


Figure 6.4: Timeline of top 3 most attacked regions monthly

6.5 Use Case 5: Ratio of children's death to monthly attacks

In addition to the results derived from the output in our integrated data, we illustrate how other humanitarian open data such as uadata[12]could be used to conduct further analysis of other war aspects in Ukraine.

Figure 6.5 presents a monthly record of attacks and children's death between April 2022 and December 2022. The first subplot reveals the monthly number of attacks, shedding light on the intensity of the situation. The second subplot highlights the tragic impact, displaying the monthly count of children who lost their lives. The third subplot captures the severity by depicting the ratio of children's deaths to attacks, offering insights into the relative impact on children. This graph serves as a valuable tool for analyzing trends and understanding the gravity of the situation that could provide a reference for humanitarian organizations.

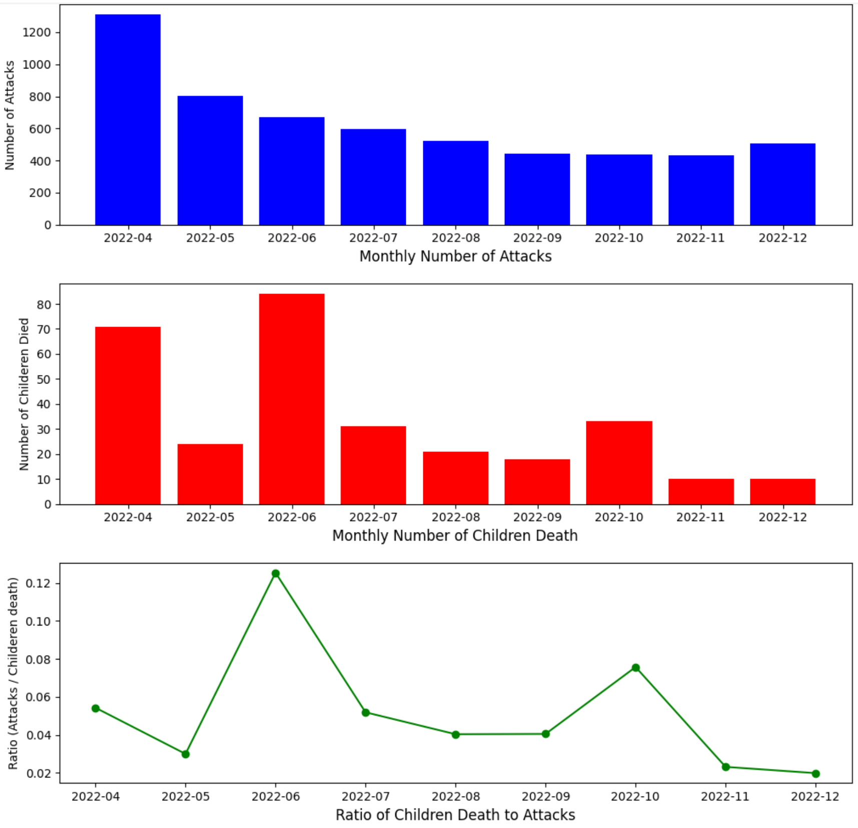


Figure 6.5: Ratio of children's death to monthly attacks

Chapter 7

Discussion and Conclusion

This thesis 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 EoR and CH. 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 links, missing links, and links to content that requires access permission. Our examination shows that 1.1% and 11.6% are such invalid links in EOR and CH, respectively. The presence of invalid links in the datasets indicates that the information gathered from social media platforms may not be reliable or complete. This problem is more significant for CH. This limit has been discussed in some previous research [4]. Further assessment and validation are required for the source and resulting datasets.

While CH is mostly about damaging events, EoR consists of some military events. A beneficial next step is to remove those events. The resulting dataset could be further enriched with information about the type of buildings, schools, etc. Moreover, the labels such as that of cities, and provinces could be enriched with multilingual information. Furthermore, The resulting datasets from this thesis project have potential for social impact, by providing valuable insights into the damage caused by the conflict and informing resilience projects. We emphasize the limitation of our data that it does not provide the exact reality of the conflict, however, we are one step closer to the truth. To prevent misuse, measures such as data anonymization, ethical guidelines, and data governance should be implemented before publishing the results openly, and therefore, the resulting data is set to private and only available upon request. Protecting personal privacy which was highlighted previously in chapter 2 paper[8] rights and promoting transparency in data collection and analysis methodologies are essential to us. By adhering to these measures, the resulting data can be harnessed effectively, maximizing its social impact while safeguarding against potential misuse.

Regarding future work, we plan to extend our dataset by integrating shelter data from external sources such as[1]. This will not only offer a more comprehensive picture of the conflict-ridden regions but will also allow us to investigate in detail areas under attack. By identifying regions lacking adequate shelter facilities in proximity to the attacked areas, we can propose new locations for shelter establishment, contributing significantly to disaster management and planning for potential emergencies. Additionally, Our approach can be extended to include additional datasets, such as ACLED [10] and WikiEvents [6] to construct a more inclusive and accurate estimation of resilience needs. And SHACL could be used for further validation.

Moreover, the graphical analysis based on use case **??** could be inspected in the future by comparing it to records provided by the UN [13] which brings attention to the tragic reality of children's deaths. It serves as a stark reminder of the profound loss and human tragedy experienced by children in war zones. The data raise an urgent need for concerted efforts to address the protection of the lives of vulnerable children. This prompts critical discussions on policy, intervention, and resource allocation to effectively highlight and prevent child mortality, fostering a global commitment to safeguarding the well-being and future of our children.

Finally, our approach along with the proposed pipeline and use cases, is not limited to the specific datasets and context of Ukraine, as it can be adapted for other types of datasets related to geo-annotated events or the resilience of different countries. This adaptability highlights the broader applicability of our methodology in addressing similar challenges worldwide.

Acknowledgement

I sincerely appreciate my supervisor Shuai Wang, for his constant guidance and support throughout this thesis project. Shuai's dedication to mentoring and nurturing my academic growth has truly encouraged me and pushed me beyond my limits, helping me discover my full potential. I also express my gratitude to Ronald Siebes and Eirik Kultorp for their valuable assistance, feedback, and insightful support to facilitate access to essential resources that substantially contributed to the realization of this thesis project.

Bibliography

[1] shelter-kyiv regional military administration. https://koda.gov.ua/en/the-public/open-data/shelter/, Jan 2022. Accessed on 15 July 2023.

[2] Dmytro Chumachenko and Tetyana Chumachenko. Ukraine war: The humanitarian crisis in kharkiv. *BMJ*, 376, 2022.

[3] (CIR). CENTRE FOR INFORMATION RESILIENCE. https://www.info-res.org/, 2020. Accessed on 30-4-2023.

[4] Kate Crawford and Megan Finn. The limits of crisis data: analytical and ethical challenges of using social and mobile data to understand disasters. *GeoJournal*, 80:491–502, 2015.

[5] Paola Espinoza-Arias, María Poveda-Villalón, Raúl García-Castro, and Oscar Corcho. Ontological representation of smart city data: From devices to cities. *Applied Sciences*, 9(1), 2019.

[6] Vasilis Kopsachilis, Nikos Vachtsavanis, and Michail Vaitis. 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, 2023.

[7] Angelie Kraft and Ricardo Usbeck. The ethical risks of analyzing crisis events on social media with machine learning. *arXiv preprint arXiv:2210.03352*, 2022.

[8] Angelie Kraft and Ricardo Usbeck. The ethical risks of analyzing crisis events on social media with machine learning. *arXiv preprint arXiv:2210.03352*, 2022.

[9] leedrake5. leedrake5-Russia-Ukraine. GitHub Repository, 2022. Accessed on 11 July 2023.

[10] Clionadh Raleigh, rew Linke, Håvard Hegre, and Joakim Karlsen. Introducing acled: An armed conflict location and event dataset. *Journal of peace research*, 47(5):651–660, 2010.

[11] John Smith. The role of multilingualism in promoting resilience and knowledge sharing. *Journal of Global Resilience*, 15(3):123–145, 2022.

[12] uadata.net. uadata. https://github.com/uadata/uadata/blob/main/zlochiny-rf.csv, Year. Accessed on 11 July 2023.

[13] UN News. Child deaths hit "tragic milestone" in ukraine. UN News, April 3 2023. Accessed on 11 July 2023.

[14] Willem Robert Van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. Design and use of the simple event model (sem). *Journal of Web Semantics*, 9(2):128–136, 2011.

[15] veelenga. Github - veelenga/rblist: Russia ban list, 2023.

[16] Marcio Victorino, Maristela Terto de Holanda, Edison Ishikawa, Edgard Costa Oliveira, and Sammohan Chhetri. 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):443, 2018.

[17] vshymanskyy. Github - vshymanskyy/standwithukraine: standwithukraine banner and related documents, Jun 2023.