From real world mining incidents to descriptive data visualisations by developing a Knowledge Graph

Laurens Beck, 2714031

Vrije Universiteit Amsterdam, L.e.f.beck@student.vu.nl

Abstract. Within the CARPA project, a forum and crowdsourcing application are developed to gather mining-related incidents to gain insights about the incidents. The incidents are submitted to the forum by mineworkers and NGOs. Currently, most submitted incidents are from the Democratic Republic of Congo (DRC). The incidents vary from deaths at mining sites to corruption cases. Unfortunately, most submitted incidents do not contain enough information, which makes further processing difficult. This research investigates the submitted incident data and makes an effort to enrich the data in order to achieve visualisation and analysis purposes for the CARPA stakeholders by developing a Knowledge Graph. To accomplish this, data enrichment requirements and Knowledge Graph construction methodologies according to best practices are investigated. By combining various datasets with the incident data, the Knowledge Graph is developed with a corresponding ontology in an iterative design science process. A Knowledge Graph facilitates the users of prompt data connections without manual data decryption. Furthermore, the Knowledge Graph is evaluated using competency questions and a visual dashboard where the incidents are plot on a map. In addition, a Natural Language Processing (NLP) script is developed as a Proof of Concept that can automatically categorise the incident types and calculates when possible coordinates.

Keywords: Mining Incidents · DRC · Knowledge Graph · Ontology · RDF triples · Design Science · Data visualisation

1 Introduction

The research project "From real world mining incidents to descriptive data visualisations by developing a Knowledge Graph" is carried out as a Master's thesis project in the field of Information Science at the Vrije Universiteit of Amsterdam (VU) [1]. This project is a contribution to the Crowdsourcing App for Responsible Production in Africa (CARPA) project [2], led by researchers of the Universiteit van Amsterdam (UvA) [3] in collaboration with the Pole Institute [4] located at Goma in the Democratic Republic of Congo (DRC).

1.1 Motivation

In 2016, the United Nations included Responsible value chains in the Sustainable Development Goals (SDGs) for 2030 [9]. This highlights the international urgency to ensure that the products used on a daily basis, such as our smartphones, are produced responsibly with a fair value chain. Therefore, companies operating in low governance countries are encouraged to produce their products responsibly. Unfortunately, the working circumstances are often poor at mining sites in the DRC. Mining companies are the only sources of income [21] which gives workers no choice but to work here, no matter the working circumstances. This is contradictory with the United Nations

[1] https:/vu.nl
[2] https:/carpa.io
[3] https:/uva.nl
[4] https://www.pole-institute.org/

goal for sustainable value chains and should therefore be improved. The CARPA project aims to contribute to this. Specifically, the goal of the CARPA project is to support responsible production in mines within various countries in Africa. The mines of interest for this research, mine various raw materials such as cobalt, gold, tin, and copper and lay within the DRC. The raw materials are used by companies around the globe with various end products such as batteries and construction work.

The CARPA project and its accompanying application aim to encourage the dialogue within communities and stakeholders of development sights, with the goals to create solutions will be created for incidents that occur in mines. In the CARPA project, incidents vary from reporting of corruption to gas explosions. The initial stakeholders are NGOs's, local government organisations, and local community leaders. In the future, this may expand. The CARPA project does not aim to provide a solution for the incidents, rather provide a system in which local stakeholders could start dialogues and discussions to improve the circumstances. [6].

1.2 Problem definition

The users of the CARPA application are mostly miners and NGOs in the Democratic Republic of Congo. Currently, the reporting of incidents by the users is unstructured and incomplete. This implies the following: the application consists of a form that includes two tabs, in which the only mandatory information that must be filled out is an incident title and a description. The incident type, date, location, and additional attachments are not mandatory. In the first iteration cycle, researchers of the project have identified that users in the experiment group could not provide this information each time they filled out the form. To decrease complexity, this was made not mandatory anymore[6].

Currently, submitted incidents are manually converted to structured data manually by a domain expert who contributes to the CARPA project. This is very labour inefficient and therefore rather expensive, as it may cost days to complete all the information. In many incident cases, it is not even possible to fill all corresponding attributes of an incident. The goal of this thesis is to create a Knowledge Graph (KG) that contributes to the enrichment of the incomplete data. The argued reason for a KG is elaborated in Section 1.4 Scientific and practical contribution. The KG leads to visualisations of incidents, eventually enabling a dashboard for the initial stakeholders with a descriptive purpose to explore the incident data.

1.3 Research questions

In order to make the CARPA incident data more complete, a KG will be created for the incomplete incidents submitted by the users of the CARPA application. Furthermore, visualisations may be developed of the incidents to enable data exploration and analysis. This research answers the following question:

- *How can the existing incomplete CARPA incident data be enriched so that it can benefit visualisation and analysis purposes?*

In order to answer this question, it must be determined what data is available from the CARPA application. The data is analysed in its database structure to confirm whether it contains useful metadata. Therefore, the following subquestions have been formulated:

- *SQ1: What requirements should be considered for the enriched data to create a knowledge graph in this project context?*
- *SQ2: How can the knowledge graph best be constructed according to best practices?*
- *SQ3: How can the knowledge graph be visualized in such a way that it is useful for CARPA stakeholders?*

SQ1 is intended to explore a knowledge graph solution and cover initial conversations with stakeholders and domain experts within the CARPA team. The scientific contribution of the knowledge graph is largely within the evaluation method, which is covered by SQ2. SQ3 covers the question about potential expansion of the CARPA project, for instance to other countries.

1.4 Scientific and practical contribution

The CARPA project was introduced by Bwana, et al. in 2020 [6]. That study highlights the iterations the application has gone through so far, explaining various design decisions made in the project. Currently, the reported incidents cannot be managed properly, and previously submitted incidents are analysed manually to enrich the data. However, this is not maintainable when the CARPA project would like to scale up in the future. Furthermore, there is a desire of visualisations for the incidents and this is currently not feasible due to the current incident data processing. The KG, is the first step to accomplish structure. Therefore, this project will open a door for various new opportunities. A KG facilitates the users of prompt data connections without manual data decryption. One could argue that a similar goal could be achieved using MySQL or another database environment, but the KG combines the database schema and the database in the same place where database structures separate these. This makes a KG more flexible and allowing context. Moreover, the semantic features of the graph make it more accessible for the science community.

The research investigates reusable methods for enrichment of unstructured incident reports to structured data. For this thesis, a design science research is performed, which is further discussed in section 3. Research strategy & research methods. Hevner et al., suggested in 2004 three types of research contributions within design science research [12]. The first contribution is described as an artefact which is a practical contribution. For this project, the knowledge graph could act as this artefact. Another benefit is the extension of heterogeneous data in KGs above pure database structures. KGs capture the environment of individual entities better than separated individual entities in widely spread, used database contexts [7].
The next type is the addition to the knowledge base, which is described as "The creative development of novel, appropriately evaluated constructs, models, methods, or instantiations that extend and improve the existing foundations in the design-science knowledge base are also important contributions"[12]. The choices made and rules applied for the creation of the KG will form an ontology for the incidents, explaining the relationships between the objects of interests within the domain. Furthermore, the evaluation of the KG is of scientific value to the science community. Lastly, the evaluation of the KG will assess the graph that answers to the overall usability.

2 Related Work

In a recent study, Fensel et al. explore the definition of a KG and state that "Knowledge Graphs are very large semantic nets that integrate various and heterogeneous information sources to represent knowledge about certain domains of discourse" [10]. For this project, the scale of the semantic net is rather small, since that dataset is not large. However, if a KG is designed well, it can be more easily expanded if upscaling the application is desired. The KG for the CARPA project is constructed from crowdsourced data. A project in the Netherlands uses a similar principle regarding crowdsourced incident management, namely WhatsApp Buurt Preventie [5]. This system provides a platform for neighbourhoods to report suspicious behaviour. All incidents are reported to the police, who manage the notifications. This project is interesting for the data categorisation.

Most large internet services used in our daily life are conceptually KGs. Google coined the term in 2012 [10]. Since then, many large tech companies built their service infrastructure upon KGs or similar structures. For example, Wikipedia is one of the richest knowledge bases on the web, with many contributors around the globe. The Wikimedia group [6], the overseeing foundation that runs Wikipedia, manages Wikidata as well. Wikidata is fundamentally different from Wikipedia by representing knowledge in a graph structured manner instead of free text [23]. This structured format enables users to query all data using a SPARQL interface [7].

Now, the knowledge graph in this research is a domain-specific knowledge graph regarding mines in the Democratic Republic of Congo. Kejriwal has written a book about the construction of domain-specific knowledge graphs [14]. Furthermore, a recent study conducted a survey among 140 specific KG constructions within seven domains [1]. This survey is of great benefit for the research to examine other domain specific approaches, including their evaluation techniques. Additionally,

[5] https://www.wabp.nl/
[6] https://commons.wikimedia.org/wiki/Main$_{P}age$
[7] https://query.wikidata.org

in a recent study, researchers developed a Dutch Maritime History KG. The development process is explained in detail, which provides useful insights in the design process of a KG [20].

Researchers state that HTML has boundaries when it comes to accessing structured data operating underneath web pages. XML, RDF, RDFS and OWL are various standards that were developed for syntaxes and data models. These were all developed having the vision of covering all kinds of intelligent processing of data on the Web in mind. [11] Schema.org [8] is an initiative of the major search engines to improve that situation. Their goal is providing a single schema across topics differing from flight tickets to dinner reservations. Approximately twelve million sites use Schema.org. The driving factor for Schema.org was to make it easier for webmasters to publish their data, since data types contain predefined definitions via the Schema.org methodology. Various design decisions had to be made in order to create the success. Some objectives with which the developers had to deal are decisions in syntax, polymorphism, entity references variate in many models and incremental complexity (a good balance between simplicity and being able to create more advanced applications). Key lessons from Schema.org are threefold. Firstly, make it easy for publishers/developers to participate. Secondly, make the specifications a compact as possible, since no one reads long specifications. Lastly, complexity has to be added incrementally, over time.

For the KG to be successful, good information is crucial. One of the main challenges of the data enrichment is filling the blanks with data, as shown in Fig. 1. According to Nadeau, Named Entity Recognition (NER) could significantly increase the usefulness of the data [15]. By processing the incident descriptions in the CARPA dataset, key features as for instance organisations and locations can be extracted to fill the blanks.

3 Research strategy & research methods

For this thesis, A design research and action research are both plausible options. In literature, these two methods are mostly treated separately. However, there are some convergences as well [8]. Design science is a research methodology in the interoperative research paradigm. Based upon the requirements of the stakeholders who desire a KG, it seems to be the most suitable approach. Within a design science research, one tries to iteratively solve a problem within a community of practice. As stated in section 1.4 Scientific and practical contribution, an artefact (KG) will be developed. Besides this thesis, the whole project consists of four artefacts. The KG, the ontology, the final dashboard, and the incident categorisation script. Researchers described design science in 2014 as the "scientific study and creation of artefact as they are developed and used by people with the goal of solving practical problems of general interest" [13]. One must be aware that the creation of a KG is no research on its own. Academic qualities such as analysing, explaining, justifying, and arguing choices based upon literature must be executed. Furthermore, critical evaluation and knowledge creation are necessary. According to Hevner et al., Design-science research follows seven guidelines. These are Design as an Artefact, problem relevance, evaluation, research contributions, research rigour, design as a search process, and communication of research [12]. The first three will be elaborated here, the other four will be simultaneously carried out throughout the project. For this research, the proposed information systems research framework of Hevner will be used in combination of these steps.

Some aspects of the action research methodology can be taken into account. The action research process starts with diagnosing the problem. Then, together with the practitioners, an action plan is made based on theories. Next, the action plan is implemented and later on evaluated. If necessary, the implemented artefact is then revised using theory, which may lead to an improved version. Lastly, the gathered knowledge of the project is shared with the stakeholders and optionally with the science community [3].

3.1 Designing an artefact

Data collection & preparation The domain expert in the CARPA team manages the incidents coming in and manually enriches the data of basic background information. This background

[8] https://schema.org

information involves mining site, location, country, incident type, local name, and company name. Fig. 1 contains a sample of the data. It is very challenging since the finished product is left with many blanks, which makes the ontology building process more complex. Moreover, local company names change frequently, which makes it rather difficult to trace the path from the mining site to the initial investors. One mining site could have three different names due to changing joint ventures. Therefore, maintainability of the data is where the challenge lies. Currently, the research methodology consists of basic searches on the Internet, in some cases there are reports attached, and applying domain experience. The incidents are reported and posted on the CARPA application. The semi-structured interview with the domain expert is described in Appendix A. His choices and decisions are of great contribution for the data links that will be applied in the KG.

Case name	Country Nan	Mining site	Location	Intl Company name	Local entity name	Case type	Incident/initiative
Banro ne remplit pas ses obligations	DRC		Luhwindja	Banro	Twangiza mining	Community Development	Incident
Violation du droit à l'indemnisation en cas de délocalisation	DRC		Luhwindja	Banro	Twangiza mining	Community Development	Incident
Frontier Service Group, fait partie d'une joint venture de raffinerie d'or, qui a des avec la	DRC		Bukavu	Frontier Services Group	Congo Gold Raffinerie	Corruption	Incident
Violation des droit de l'homme	DRC	Misisi	Bukavu			Corruption	Incident
Paiement taxes groupe armé Nyatura site minier Bihovu	DRC	Bihovu; Lumbishi	Kalehe			Corruption	Incident
Gécamines: rapport sur des pratiques peu transparentes	DRC		Lubumbashi		Gecamines	Corruption	Incident
Tracasseries à Muderi	DRC					Corruption	Incident

Fig. 1: Sample of CARPA data

For this research, there is no time to gather field experience and built a network. Therefore, there must be relied on reports and basic internet searches when new incidents occur during the research phase. According to the KG definition, various information sources are used to construct a KG [10] [18]. For this project, this is certainly relevant duo to the small size of the dataset. Therefore, other data must be found to enrich the dataset. Potentially, this could include information about certain mines.

Develop an ontology Researchers propose various reasons why one would develop an ontology [17]. Firstly, the common understanding of information structures among people or software agents is desired. Other reasons are the reusability of domain knowledge, separating domain knowledge from operational knowledge, making domain assumptions explicit, and lastly, analysing domain knowledge.

One might argue that ontology development and object-oriented design are similar. However, design decisions with in object-oriented programming are made upon operational properties of a class, and ontology development uses the structural properties to make decisions. Within an ontology, classes are used to describe concepts of a domain. These classes could have subclasses so that concepts are explained more explicitly. For instance, a mine in Africa is a class of mines. The mine can be categorised by the raw materials it delves, such as a gold mine, which is the subclass. The location of the mine is a property of the class mine. The general rules for creating an ontology are: defining classes; arranging classes in a taxonomic hierarchy; defining slots; and filling in values for slots of instances.

Researchers suggest three different approaches when it comes to developing a class hierarchy. Firstly, a top-down approach uses the definitions of the most general concepts within a domain. Next, is the bottom-up approach, which starts with the granularity of a domain and then work up. The last approach is a combination of top-down and bottom-up [22], also named the hybrid approach [17]. After the interview with the domain expert, for this project the most suitable approach will be the combination due to the many gaps in the existing data (See Appendix A). Ultimately, existing ontologies can be used, such as [9]. However, the main part must be developed from scratch.

Pre-annotators to acquaint the first patterns in the data are necessary. Furthermore, a dictionary can be created so that specific words can be tagged as part of a certain entity. Lastly, rules can be applied that can classify words in specific circumstances. Tools as WebProtogé [10], Web VOWL [11] could be used to explore the data and try to create ontologies. W3C developed RDF, which is a language for encoding knowledge on the Web. The framework standardises knowledge formats so that it is searchable for Web agents [5].

[9] schema.org
[10] https://webprotege.stanford.edu/
[11] http://vowl.visualdataweb.org/webvowl.html

3.2 Problem relevance and KG set up

The KG is constructed in the problem relevance phase. This phase has some overlap with the designing phase, since it is a vivid process. One must go back and forth to the drawing table whilst constructing. An option for creating the KG is GraphDB which can help with indexing with the purpose of semantic research whilst performing text analysis [12]. Within GraphbDB, data conversion can be directly executed, as well as SPARQL queries to load the data. An advantage of GraphDB is its friendly user interface. Additionally, it is possible to import and extract RDF mappings to different repositories or programs. This is an advantage for the usability and interoperability. Furthermore, the graph environment makes the tool overall sufficient for our purposes. Researchers have conducted a study in which they compared the most common knowledge graph frameworks and discuss evaluation methods for each [18].

The data will be loaded in GraphDB and the first RDF triples will be constructed. These triples are interpreted by GraphDB to create the graph. A triple consists of a subject, a predicate, and an object. A triple could look like this:

$$< carpa : mine >< schema : has_location >< carpa : location > \tag{1}$$

This triple tells the graph that a certain mine has a certain location. The created ontology will be used to construct the triples together with other ontologies like schema.org and the RDF standard. When the ontology is finished, it is loaded into GraphDB so that the KG uses the developed ontology.

3.3 Design evaluation

The third guideline of Hevner et al. is design evaluation. This consists of rigorous evaluation of the artefact using well-executed design methods [12]. The evaluation methods variate from observational, analytical, experimental, testing, and descriptive.

Paulheim stated in 2016 that there is no silver standard when it comes to evaluation methods of KGs and their corresponding ontology data model, since most approaches are only evaluate one domain specific KG [18]. The most suitable approach tends to be descriptive evaluation, where competency questions help evaluate the utility of the graph with SPARQL queries [24][4]. The competency questions are developed together with the CARPA domain experts and follow the guidelines of Azzaoui et al. [2]. The team of domain experts consist of developers and researchers with 15 years of experience in the field in various African countries. These domain experts represent the stakeholders of the project. Therefore, they are able to desing the competency questions. The competency questions are divided in categories as shown in Table 1. In the evaluation phase, the competency questions are answered using SPARQL queries in the GraphDB environment and within the dashboard. Together with the CARPA team, it was decided that the visualisation of the data as an end product can contribute to the evaluation method to test the usability of the graph.

Paulheim held a survey of approaches and evaluation methods for domain specific knowledge graphs [18]. This study conducts mostly refinement methods for existing knowledge graphs and distinguishes seven different evaluation metrics. These metrics are Hits@N, MRR, MR, Accuracy, Precision, Recall, and F-measure. Each metric has a different function with different purposes. The study concludes that there is no standard that can be used in each case, since most knowledge graphs created serve one single and unique goal [18]. However, Chen et al., propose in a recent study a framework for exactly this kind of structures [7]. By selecting certain quality dimensions with corresponding metrics, their approach evaluates these metrics on feasibility and scalability. The researchers conclude that this is necessary to determine the "fitness of purposes" of the KG. What effect the framework has in practice, is not yet discussed and is left for future work.

The most suitable approach to evaluate the dashboard is by Black Box testing [12] since the dashboard is an artefact that will have users. Black box testing is part of acceptance testing where a user interacts with the environment in the analysis phase [16]. By giving users tasks to complete within the environment, failures, and defects of the environment can be detected whilst

[12] https://www.ontotext.com/products/graphdb/

the researcher observes the users' interaction behaviour. The tasks consist of the competency questions developed by the CARPA team domain experts.

Table 1: Competency Questions

Category	CQ
1. Mineral	1.1 How many incidents are related to a specific type of mineral? 1.2 What violation type is most common with extracting a particular type of mineral
2. Time	2.1 How many incidents have been raised relating to this company over a period of time? 2.2 How many incidents have been raised related to the extraction of a specific mineral over a period of time? 2.3 How many incidents have occurred over a defined period of time? 2.4 What is the comparison of incident numbers between one period and another?
3. Location	3.1 How many incidents are related to a mineral type within a region? 3.2 How many incidents in a region are of a particular violation type? 3.3 How are incidents spread between regions or within regions? 3.4 How many incidents were reported in the region of Nord-Kivu in 2020?
4. Company	4.1 How many incidents are related to a particular company? 4.2 Which violation type is most often reported regarding a particular company?
5. Armed Group	5.1 What armed groups are present at mines where incidents occurred?
6. Labour	6.1 Is there child labour at mines where incidents occur? 6.2 How many people work at mines where incidents occur?

4 Results

In this chapter, the results of the research are elaborated. Since the research method of this thesis is design science, the sections are in line with this approach. Firstly, the first iteration of the design process is explained, starting with the data collection and preparation of the data. Followed by the data ontology modelling.

4.1 First iteration

Data collection & preparation Firstly, all data available from the CARPA project was collected, and research efforts were made to find other datasets which could potentially enrich the data. The research resulted in the collection of two other datasets, which are integrated in the KG. First, a dataset from the IPIS research institute, which kept track of artisanal mining sites in Eastern DRC between 2009 and 2021. They visited various mines and collected over 80 different attributes per visit. The dataset was extracted from africaopendata.org [13] and contains over 5000 visits. After cleaning and decision-making which columns to use, a useful dataset was left. A sample of the data is illustrated in Fig. 2.

latitude	longitude	id	pcode	mine	visit_date	province	province_old	territoire	collectivite	groupement	village	workers_numb	mineral	selling_points_	final_destinatic	mineral2	armed_group	type_armed_group	childunder15	women
0.33188	28.71258	130	codmine00191	Eohe	2009-01-01T02:00:00Z	Nord-Kivu	Nord-Kivu	Lubero	Bapere	Baredje	Maendeleo	300	Gold					0		
0.32153	28.69916	101	codmine00192	Eita	2009-01-01T02:00:00Z	Nord-Kivu	Nord-Kivu	Lubero	Bapere	Baredje	Tembe	110	Gold					0		
0.54499175	28.1851423	166	codmine00242	Mungu Iko	2009-01-01T02:00:00Z	Nord-Kivu	Nord-Kivu	Lubero	Bapere	Bapaitumba	Ehoto		Gold			Coltan	FARDC	FARDC - Pas de données sur les ingérences		
-0.352529	28.884528	163	codmine00260	Kiviri/Tayna	2009-01-01T02:00:00Z	Nord-Kivu	Nord-Kivu	Lubero	Bamate	Luenge	Vunyakavuliwa		Gold			Diamant	FDLR	Groupe armé étranger		
-0.036707	28.903945	172	codmine00272	Makanga	2009-01-01T02:00:00Z	Nord-Kivu	Nord-Kivu	Lubero	Batangi	Musindi	Ngohi		Gold			Diamant	FDLR	Groupe armé étranger		
0.88749167	29.422085	70	codmine00286	Kasongo	2009-01-01T02:00:00Z	Nord-Kivu	Nord-Kivu	Beni	Beni-Mbau	Batangi-Mbau	Mamove	50	Gold					0		
0.595655	29.24183	30	codmine00288	Mununze	2009-01-01T02:00:00Z	Nord-Kivu	Nord-Kivu	Beni	Beni-Mbau	Baswagha-Madiwe	Mununze	20	Gold					0		
0.459223	29.09466	131	codmine00290	Kiboto	2009-01-01T02:00:00Z	Nord-Kivu	Nord-Kivu	Beni	Beni-Mbau	Baswagha-Madiwe	Cantine/Aloya	300	Gold					0		
-1.696243	28.884336	38	codmine00307	Ruzirantaka (rivière)	2009-01-01T02:00:00Z	Nord-Kivu	Nord-Kivu	Masisi	Bahunde	Muvunyi-Shanga	Ruzirantaka	27	Cassitérite			Tourmaline	FARDC	FARDC - Pas de données sur les ingérences		
-1.584475	28.893558	148	codmine00309	D3 Bibatama	2009-01-01T02:00:00Z	Nord-Kivu	Nord-Kivu	Masisi	Bahunde	Muvunyi-Karuba	Rugeshi/Humule	1150	Cassitérite			Coltan	FARDC	FARDC - Pas de données sur les ingérences		

Fig. 2: IPIS data Sample

[13] https://africaopendata.org/dataset/artisanal-mining-site-visits-in-eastern-drc

The other dataset was retrieved from Wikidata. By executing a SPARQL query in the Wikidata Query Service [14], a dataset was created with all mines in DRC available on Wikidata, including their attributes such as coordinates. The optional columns included were the product, the GeoNames tag, operator, and the region. This left a set with 96 mines. The SPARQL query is shown in Fig. 3.

```
#defaultView:Table
SELECT ?mine ?mineLabel ?coords ?productLabel ?geonamesLabel ?operatorLabel ?regionLabel
WHERE {
  ?mine wdt:P31 wd:Q820477;
  wdt:P625 ?coords;
  wdt:P17 wd:Q974.
  OPTIONAL { ?mine wdt:P1056 ?product. }
  OPTIONAL { ?mine wdt:P1566 ?geonames. }
  OPTIONAL { ?mine wdt:P137 ?operator. }
    OPTIONAL { ?mine wdt:P131 ?region. }
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
```

Fig. 3: Wikidata query

After the retrieval of the DRC mines from Wikidata, the label tags were cleared from the data and all columns were transformed to lower cases within GraphDB using GREL functions. Furthermore, the coordinates' column was split to a longitude and latitude column. the result of this cleaning are illustrated in Fig. 4.

wikidata_ref	mine	latitude	longitude	product	geonames	operator	region
http://www.wikidata.org/entity/Q22380285	Makongo	-1.739722222	28.198333333		8435412		North Kivu
http://www.wikidata.org/entity/Q3541532	Shituru	-11.000681	26.756945	copper	922161		
http://www.wikidata.org/entity/Q14157452	Etoile Mine	-11.632134	27.580168	copper			
http://www.wikidata.org/entity/Q15100601	Kipushi mine	-11.769444444	27.235555555	copper	922267		Kipushi
http://www.wikidata.org/entity/Q15100601	Kipushi mine	-11.766111111	27.236666666	copper	922267		Kipushi
http://www.wikidata.org/entity/Q15100601	Kipushi mine	-11.769444444	27.235555555	zinc	922267		Kipushi
http://www.wikidata.org/entity/Q15100601	Kipushi mine	-11.766111111	27.236666666	zinc	922267		Kipushi
http://www.wikidata.org/entity/Q15100618	Kananga Mine	-10.666512	25.466394	copper			
http://www.wikidata.org/entity/Q15233178	Kambove mines	-10.812868	26.585745	copper	923059	Gécamines	
http://www.wikidata.org/entity/Q15233178	Kambove mines	-10.8	26.583333333	copper	923059	Gécamines	

Fig. 4: Wikidata data sample

Data & ontology modelling Protégé [15] was used to create an ontology for the KG. Noy and McGuinness propose guidelines for ontology building, which is used to construct the ontology [17]. A combination of a top-down and bottom-up approach [22] were performed.

The ontology consisted of eight main classes with various corresponding subclasses. The class hierarchy is illustrated in Fig. 5. The classes are based upon the data available from all three datasets. However, the class NGO is not yet included but was anticipated for the future since the CARPA team wants to collect this data as well. In Fig. 5 all blue circles represent classes and subclasses. Each green component is a data property with a corresponding data type. The relations between the classes are written in the blue rectangles. The design decisions of this structure were closely monitored with the CARPA team.

Now that the first draft of the ontology is finished, it is exported to GraphDB so that it can interact with the data instances. Each of the three datasets are mapped using RDF in the OntoRefine [16] tool of GraphDB. The instances consist of IRIs and literals. The data are being intertwined, which gives us the result we are looking for.

In Fig. 6, the Numbi mine illustrates integrated properties and attributes. In the black square are three incidents visible, which origin from the CARPA dataset. The enrichment via the IPIS

[14] https://query.wikidata.org/
[15] https://protegewiki.stanford.edu/wiki/Protege$_{O}ntology_{L}ibrary$
[16] https://graphdb.ontotext.com/documentation/free/loading-data-using-ontorefine.html

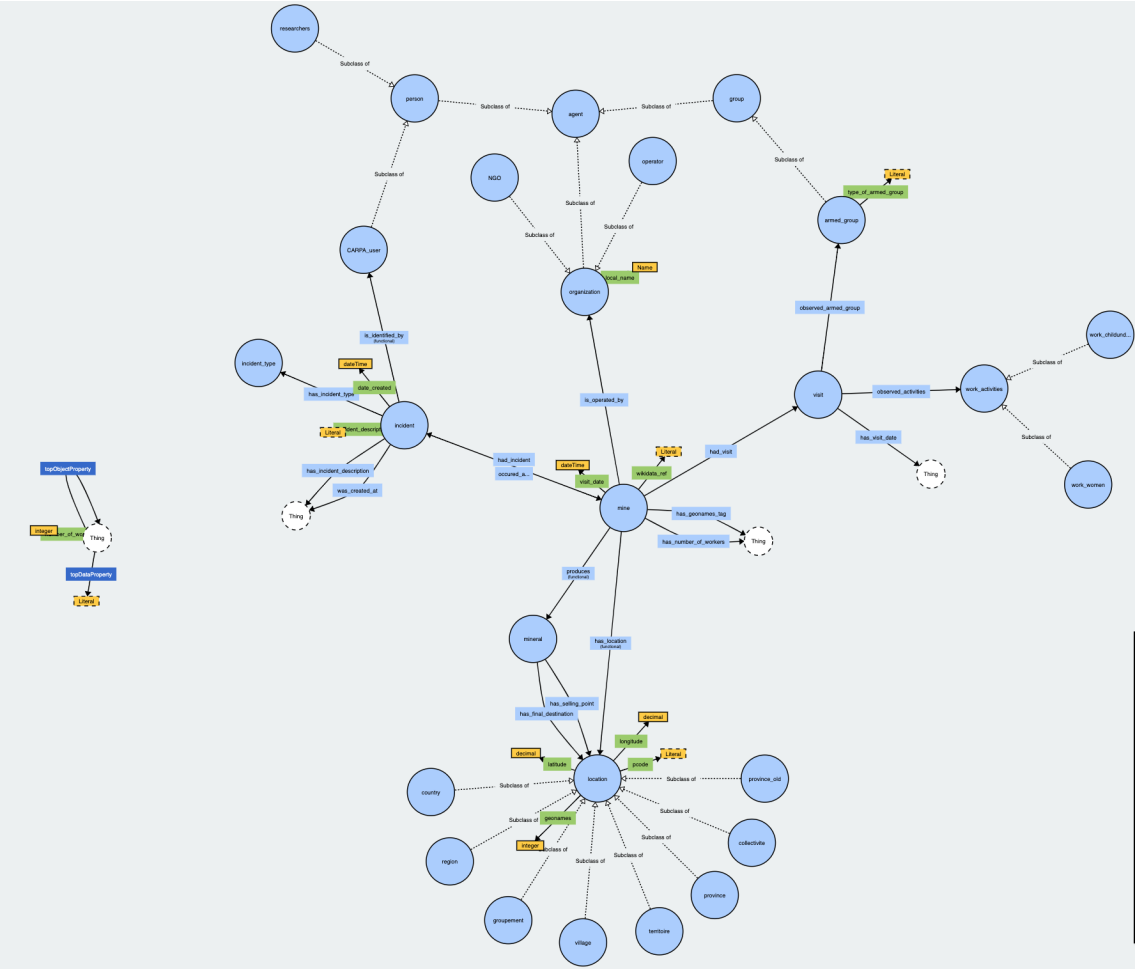


Fig. 5: Class hierarchy of the ontology

dataset is visible in the blue squares. The FARDC was present at the mine in 2009, when researchers of the IPIS institute visited the mine. Furthermore, the mine produces cassiterite and is not only a mine but functions as a distribution location for gold and mangarése (manganese) as well.

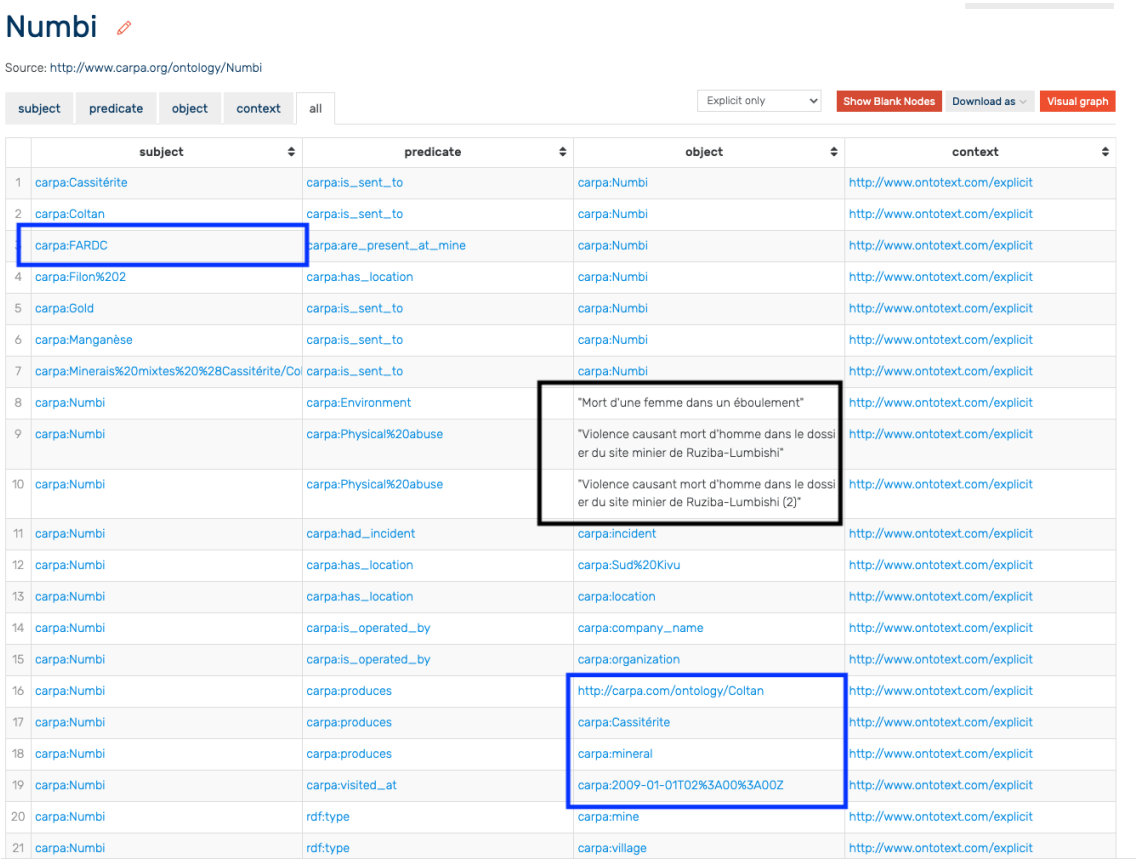
4.2 Evaluation of first iteration

Whilst modelling, it became clear that the Wikidata dataset did not contain much overlap with the other two datasets. However, since there are only 118 incidents so far, it is not peculiar that there is no overlap yet. Furthermore, whilst the CARPA project scales up, and more incidents will be submitted, the chance of the Wikidata set having overlap with the CARPA dataset will rise. Therefore, the dataset is still valuable to the project.

Concerning the KG and the ontology, the following things became clear. The IPIS dataset contains 5251 visits to 2787 different mines. The information of a single visit links directly to a particular mine. This gives a clear overview when a mine has one visit, since all additional attributes such as “work childUnder15” and ”workersNumb” clearly corresponds with that one visit. However, when a mine various visits, it is not clear which attribute corresponds to which visit. For instance, a certain mine has four visits and the “worker number” attribute was only tracked in three visits. Therefore, it is not clear which visit did not contain this number.

In addition, the incidents have the same problem. The “incident titles” stand directly between the attributes of the mine, with the “incident type” as predicate. Again, as with the visits, this led to the confusion of data attributes when multiple incidents occurred at the same mine.

Lastly, the ontology is designed so that an incident → occurred at → mine, since the location attribute was not specific enough in some cases (Sud-Kivu, which is a province, was the location value for some incidents). Additionally, The mine attribute only contained data in 35 incidents. Therefore, analysis via SPARQL could only are not complete since there are 118 incidents.



Numbi

Source: http://www.carpa.org/ontology/Numbi

subject | predicate | object | context | all

Explicit only | Show Blank Nodes | Download as | Visual graph

	subject	predicate	object	context
1	carpa:Cassitérite	carpa:is_sent_to	carpa:Numbi	http://www.ontotext.com/explicit
2	carpa:Coltan	carpa:is_sent_to	carpa:Numbi	http://www.ontotext.com/explicit
3	carpa:FARDC	carpa:are_present_at_mine	carpa:Numbi	http://www.ontotext.com/explicit
4	carpa:Filon%202	carpa:has_location	carpa:Numbi	http://www.ontotext.com/explicit
5	carpa:Gold	carpa:is_sent_to	carpa:Numbi	http://www.ontotext.com/explicit
6	carpa:Manganèse	carpa:is_sent_to	carpa:Numbi	http://www.ontotext.com/explicit
7	carpa:Minerais%20mixtes%20%2BCassitérite/Co	carpa:is_sent_to	carpa:Numbi	http://www.ontotext.com/explicit
8	carpa:Numbi	carpa:Environment	"Mort d'une femme dans un éboulement"	http://www.ontotext.com/explicit
9	carpa:Numbi	carpa:Physical%20abuse	"Violence causant mort d'homme dans le dossier du site minier de Ruziba-Lumbishi"	http://www.ontotext.com/explicit
10	carpa:Numbi	carpa:Physical%20abuse	"Violence causant mort d'homme dans le dossier du site minier de Ruziba-Lumbishi (2)"	http://www.ontotext.com/explicit
11	carpa:Numbi	carpa:had_incident	carpa:incident	http://www.ontotext.com/explicit
12	carpa:Numbi	carpa:has_location	carpa:Sud%20Kivu	http://www.ontotext.com/explicit
13	carpa:Numbi	carpa:has_location	carpa:location	http://www.ontotext.com/explicit
14	carpa:Numbi	carpa:is_operated_by	carpa:company_name	http://www.ontotext.com/explicit
15	carpa:Numbi	carpa:is_operated_by	carpa:organization	http://www.ontotext.com/explicit
16	carpa:Numbi	carpa:produces	http://carpa.com/ontology/Coltan	http://www.ontotext.com/explicit
17	carpa:Numbi	carpa:produces	carpa:Cassitérite	http://www.ontotext.com/explicit
18	carpa:Numbi	carpa:produces	carpa:mineral	http://www.ontotext.com/explicit
19	carpa:Numbi	carpa:visited_at	carpa:2009-01-01T02%3A00%3A00Z	http://www.ontotext.com/explicit
20	carpa:Numbi	rdf:type	carpa:mine	http://www.ontotext.com/explicit
21	carpa:Numbi	rdf:type	carpa:village	http://www.ontotext.com/explicit

Fig. 6: Numbi mine example

4.3 Second iteration

In the second iteration, the KG improves, by troubleshooting the KG and its corresponding ontology using the outcomes of the first iteration. In the second iteration, new artefacts are developed. Firstly, a Natural Language Processing (NLP) script to pre-process incoming incidents based on the incident description. This script is a proof of concept Next, is the implementation and configuration from the KG. Lastly, a dashboard is developed to provide geographical insights to the data.

NLP script The main research question is formed in a way that the KG development contributes to the enrichment of visualisation and analysis purposes of the existing unstructured data. Here, an effort is made in analysing the incident descriptions with Named Entity Recognition. This technique can locate and classify entities in a certain text into pre-defined categories, which can be seen as the columns of a data sheet. This form of Natural Language Processing (NLP) can potentially make the handwork of the domain expert redundant, or at least simplify it. For instance, when an incident description mentions Nord-Kivu, this will automatically be categorised as a province. The dictionary of all data instances is filled by all existing unique instances within the current data. This can potentially be expanded when the CARPA team would like to operate in countries outside of Congo.

The NLP script is created in Python and uses among other the Spacey [17] library with the French and English language module. After first testing, it became clear that the English language module was more accurate than the French module, even when translating the original incident text from French to English. This is illustrated in Fig. 7.

The dashboard requires two key components to plot the markers on the map. Coordinates and incident type. The python script is able to create both based on a location name and incident description.

[17] https://spacy.io/

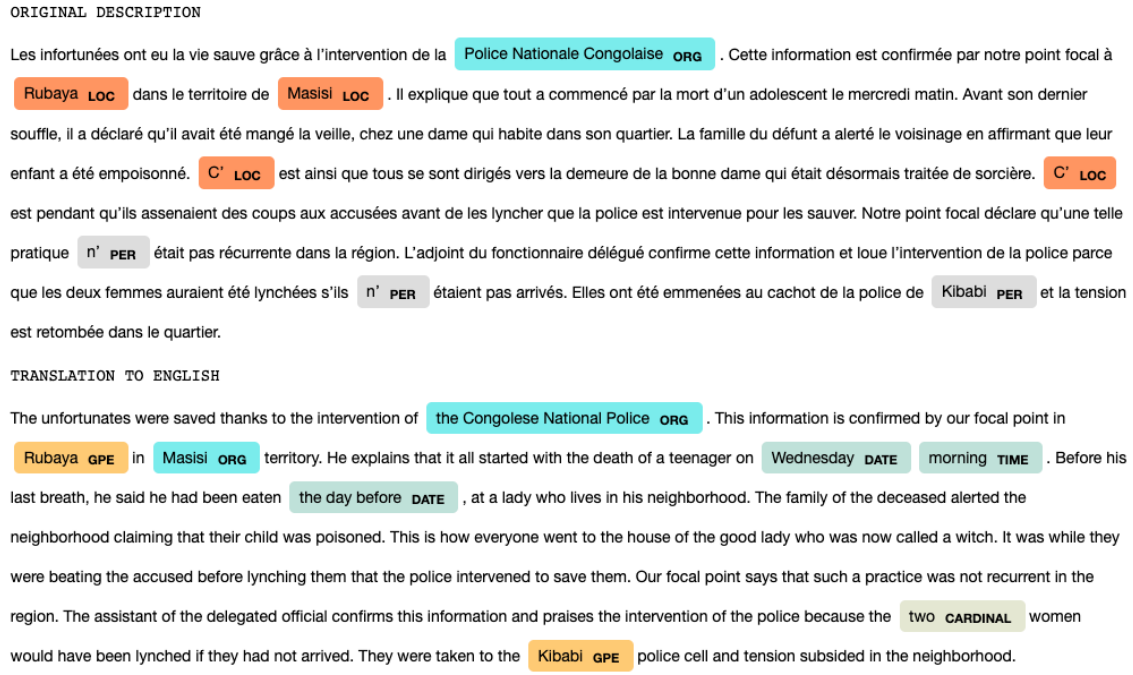


Fig. 7: Spacy French language module vs. English language module

Next, a dictionary was made manually where certain keywords are defined for incident categories. Keywords for the category physical abuse are for instance ”violence”, ”sexuelle” and ”mort”. In the case that a description includes one of these words, the incident is categorised as physical abuse. This led to eight incidents, without a category, being assigned a category. The classification is not flawless. Firstly, the keywords in the dictionary could be expanded in order to categorise more incidents. Furthermore, some incidents include words from various incident types. A decision must be made what to do in such a case.

The last part of the script is the coordinates' calculation. The reason this is done in this script instead of using the Google Maps API is simple; Google charges money to calculate the coordinates.

When the script is completely finished, all submitted incidents that have incomplete coordinates and or incident type have these values filled out as much as possible. Other values, are enriched via the KG before being visualised in the dashboard. Additionally, the coordinates are calculated upon the location column. If this column is empty in the raw incident data, the coordinates could potentially be found in the KG when a mine that exists in the KG does have coordinates. The complete python script is provided together with this thesis.

KG & Ontology Next, the ontology had to be improved, and some design decisions had to be made. Firstly, the visits of the IPIS datasets were separated from the mines. It became clear in the evaluation of the first iteration that there was confusion about data attributes concerning visits. It was misleading to which visit these attributes were linked. Therefore,”visit” became a class in the ontology with ”visit date”, ”number of”, ”work childUnder15”, ”workersNumb”, ”workWomen”, ”armedGroup” as data properties. The ”visit” receive a unique identifier starting with V.0001.

The incidents are separated from the mine as well. In the first iteration, the ”incident titles” stood directly between the attributes of the mine, with the ”incident type” as predicate. Therefore, the incident title became the unique identifier and contained all attributes corresponding to that particular incident so that there could not be any confusion. However, this assumes that incident titles are always unique. This assumption is funded since the chance of an identical incident title corresponding to the same mine is low.

All incidents that had no ”mine” value, received based upon longitude and latitude coordinates (when available) the name of the location with an ”X” marked so that it is clear that these incidents had their ”mine” value filled later. This is done so that the mines appear in the SPARQL queries.

Dashboard The dashboard has the goal to provide insights in the incident data for the users of the CARPA application. The users can be identified as for instance NGOs and researchers interested in the CARPA project. Together with the CARPA team developer, requirements for the dashboard were developed as illustrated in Table 2.

Table 2: Dashboard requirements

Requirement	Description
r1	Provide a geographical map with Africa as centre
r2	Indicate the location of the incidents on the map via markers
r3	When clicked on the marker, a summary of the incident appears
r4	Numerical indicators provide totals and other useful insights
r5	The dashboard has the ability to filter incidents based on type
r6	Geographical data can be clustered, or different incident types are categorised by colour in the first overview

Researchers held an evaluation about various Application Programmer Interfaces (APIs) regarding mapping [19]. The study compares Google Maps [18], Bing Maps [19], Nokia HERE [20], MapQuest [21], OpenStreetMap [22], Leaflet [23], Baidu Map [24], and Mapstraction [25]. The study claims that Google Maps is the most popular and could potentially meet all requirements set by the CARPA team developer. Furthermore, the way it will be used in this dashboard makes the API free of charge [26]. A platform such as OpenStreetMap requires the most coding. Therefore, it is able to offer more functionality. However, the extended functionality is not necessary for this project. Google Maps has more advanced tools to create clusters in the data, which is requirement r7. In the final analysis, all factors combined cause the choice for Google Maps.

The resulting dashboard demonstrates the geographical data mapped in Fig. 8. Currently, incident data from the KG are extracted and converted to a JSON file.

Some design decisions are made regarding the dashboard. Firstly, the attributes displayed are the minerals a mine produces, the presence of an armed group, and the number of workers. The presence of children and women are omitted, since there was only one mine where an incident occurred that included information about children or women working. In a further stage, with more data, This could be implemented. When a mine was visited numerous times, The information of the most recent visit is used. In the dashboard, the visit are data clustered under underneath the date of the last visit. The dashboard shows the locations of incidents using markers and has Africa as the centre of the map to meet the requirements r1. and r2. The summary of the incident is presented on the right when clicking on an incident meeting requirement r3. On the bottom right, indicators provide a summary of the amount of incidents within range and incident type meeting requirement r4. The filters below the map are able to filter the incidents based on type meeting requirement r5. In the first overview, the geographical data are clustered by colour. Therefore, meeting requirement r6.

4.4 Evaluation of second iteration

The completion of the second design iteration resolved the issues that arose during the first iteration. These issues are in detail elaborated in the evaluation of the first iteration. In the second iteration, the KG was finished up to a stage that the competency questions could be answered with SPARQL queries. The answers are described in table 3.

[18] https://developers.google.com/maps
[19] https://www.microsoft.com/en-us/maps/choose-your-bing-maps-api
[20] https://developer.here.com/
[21] https://developer.mapquest.com/
[22] https://wiki.openstreetmap.org/wiki/API
[23] https://leafletjs.com/reference.html
[24] https://gist.github.com/jackyliang/90b8c6cd5fd652b9d56b
[25] https://www.programmableweb.com/api/mapstraction
[26] https://developers.google.com/maps/billing-and-pricing/pricing

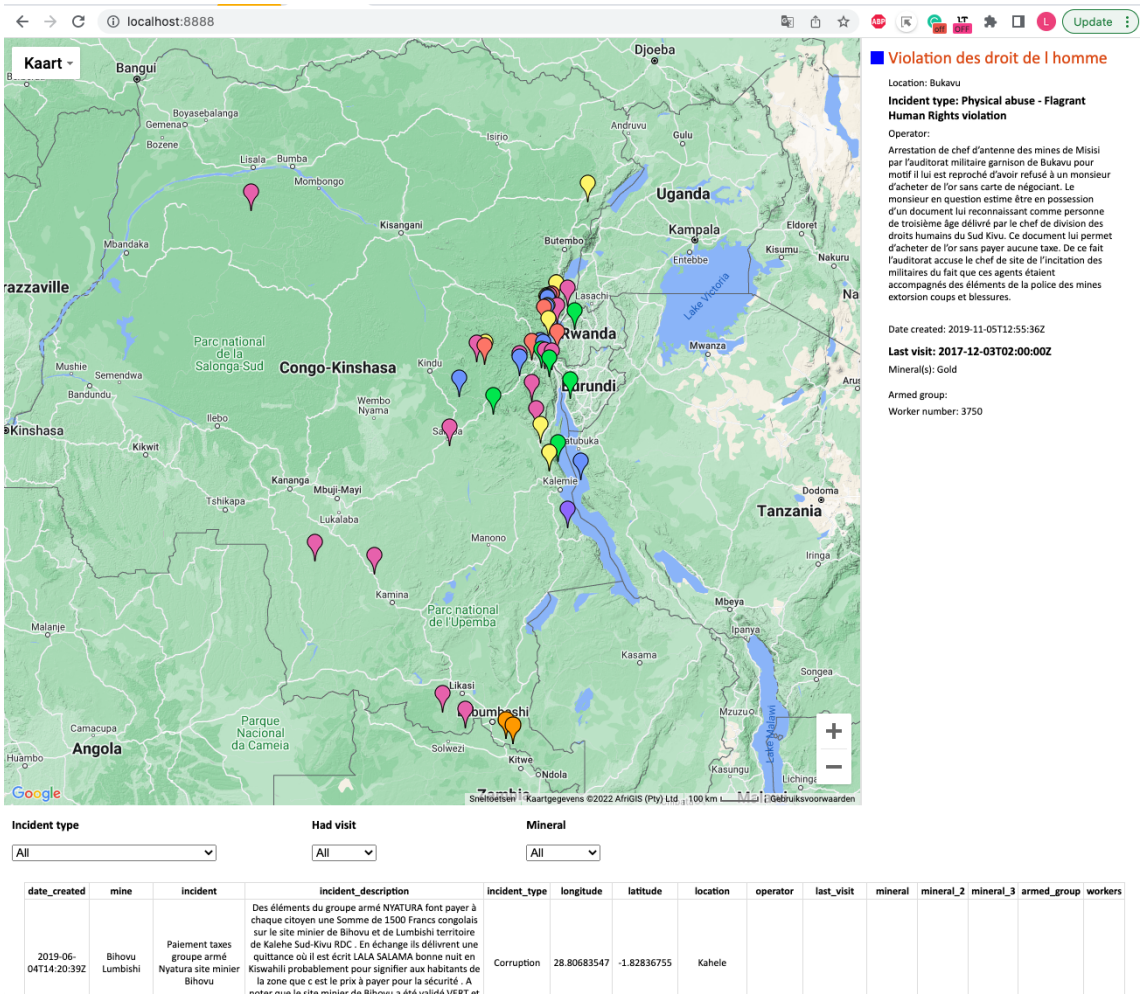


Fig. 8: Incident Dashboard

The Dashboard was developed based upon the requirements established together with a domain expert of the CARPA team and the IT developer. The dashboard is evaluated by various users who received the assignment to answer the competency questions using the dashboard. After initial testing, the competency questions that could be answered are 1.1, 1.2, 2.1, 2.2, 2.3, 2.4, 4.1, 4.2, and 5.1. This was based upon the first users not being able to solve the other competency questions, since the data to answer these questions are not included in the first version of the dashboard. Ultimately, this evaluation method tests the utility and usability of the dashboard. The black box testing method is described in chapter 3. Strategy Methods. The findings and general comments led to further improvements of the dashboard and are illustrated in Appendix B. Some findings could be classified as bugs, such as "Long descriptions extend the assigned text box" and "Clusters do not reload when filtering". When a bug was found, it was improved before the next user started testing, since bugs are in the way of usability testing. Other findings were more design decisions such as "Not clearly visible when mines are at the same location" and "Region column would be interesting to help answer the questions".

After completion of this project, the CARPA dataflow has changed. The new dataflow is illustrated in Fig. 9.

Table 3: Answers to Competency Questions

CQ	SPARQL	Dashboard
1.1 How many incidents are related to a specific type of mineral?	Query mines where incidents occurred, including produced product	X
1.2 What violation type is most common with extracting a particular type of mineral	Query mines where incidents occurred, including produced product and incident type	X
2.1 How many incidents have been raised relating to this company over a period of time?	Query mines where incidents occurred, including its operator and the date stamp of the incident	X
2.2 How many incidents have been raised related to the extraction of a specific mineral over a period of time?	Query mines where incidents occurred, including the mineral produces and the date stamp of the incident	X
2.3 How many incidents have occurred over a defined period of time?	Query mines where incidents occurred, including the date stamp of the incident, and filter on a particular time period	X
2.4 What is the comparison of incident numbers between one period and another?	Query mines where incidents occurred, including the date stamp of the incident, and filter on a particular time period twice.	X
3.1 How many incidents are related to a mineral type within a region?	Query mines where incidents occurred, including the mineral it produces and the region	
3.2 How many incidents in a region are of a particular violation type?	Query mines where incidents occurred, including the incident type and the region	
3.3 How are incidents spread between regions or within regions?	Same query as 3.2 and compare specific regions	
3.4 How many incidents were reported in the region of Nord-Kivu in 2020?	Query mines where incidents occurred, including the incident date and the region. Next, filter by Nord-Kivu and order by date.	
4.1 How many incidents are related to a particular company?	Query mines where incidents occurred, including the mine operator	X
4.2 Which violation type is most often reported regarding a particular company?	Query mines where incidents occurred, including the mine operator and the incident type	X
5.1 What armed groups are present at mines where incidents occurred?	Query mines where incidents occurred and was visited, including armed groups	X
6.1 Is there child labour at mines where incidents occurred?	Query mines where incidents occurred and was visited, including child labour	
6.2 How many people work at mines where incidents occurred?	Query mines where incidents occurred and was visited, including women	

Fig. 9: New dataflow

5 Conclusion & Discussion

In this thesis, a reusable and maintainable knowledge graph for the CARPA project regarding mining incidents in the Democratic Republic of Congo is presented. The knowledge graph was created using an iterative design and development process following the guidelines of Design Science conducted by Hevner et al. [12]. The underlying ontology was developed using three different datasets concerning mines in the Democratic Republic of Congo and resulted in more than 66 thousand RDF triples.

The main question of how the existing incomplete CARPA incident data could be enriched, is answered by enriching the data by linking the incident dataset with various datasets on the web. This is in line with related work and other KG research [20] [10] [18]. The IPIS research mine visit data was of great benefit for this research due to over five thousand visits at over twenty-five hundred mines in a time frame of 12 years in the Democratic Republic of Congo. Each visit contained many attributes that enriched half of the mines in the incident dataset.

SQ2 of how the KG is constructed according to best practices, is answered by starting the construction of the KG with data collection and preparation, where collected data was cleaned for the modelling purpose. Furthermore, the whole design process consisted of two iterations where the KG and additional artefacts were improved. The ontology was developed using a combination of a top-down and bottom-up approach conducted by Uschold & Gruninger [22]. Within the ontology design process, additional tools were used to aid the visualisation and modelling of the ontology. The ontology and the prepared data were loaded into the KG environment, where RDF triples were constructed so that the data and ontology interact. The ontology and KG are evaluated using the competency questions that are developed together with the domain expert according to best practices. Furthermore, selected competency questions are used as tasks in the black box testing of the dashboard, in order to test usability. Therefore, SQ3 is answered following the CARPA stakeholder's requirements and answering the competency questions.

The data about mines where incidents occur are enriched with data about the presence of armed militia, working activities of children and women at mines, and the mineral it produces. Moreover, geographical coordinates and location attributes such as village and province are now linked, whereas the initial data only had a single location attribute. Additionally, a dashboard was developed that includes the enriched incidents. This benefits the visualisation purposes of the CARPA Stakeholders. Finally, the developed KG offers the possibility to further analyse and explore the data. The NLP proof of concept script illustrates that new incoming incidents are enriched without manual labour, which is more time efficient and is a common practice for data enrichment. Therefore, this is the point where the pre-processing of incoming incidents will start and is the first element in the pipeline from raw incident data to visualisation.

5.1 Future work

Future work needs to be done in order to improve the KG. The goal of this thesis was making the design process from raw incident data to visualisation as transparent as possible so that improvements could be made in the future based upon our made design decisions. Linked data works best when there is a great amount of data to be linked. The KG will become increasingly more insightful when more incidents are submitted to the CARPA platform, since there is a substantial knowledge base of mines constructed that enriches the data automatically.

Currently, it is not clear whether each time an incident occurs, it is submitted by a user to the CARPA platform. Therefore, additional data input could come from the extraction of (local) news articles about mining incidents. This expands the NLP script so that it can extract the information for the KG. An example of an article that could be extracted is from mining technology [27]. The attributes from this article consist of the location (mine and province), incident description, incident type, operator, and mineral. When inserting this incident into the KG, it would be linked to a gold mine in the Katanga province operated by Glencore where at least 12 miners were killed. Next, future work should be done to evaluate the NLP script and extend it so that is more accurate, but this is for now outside my project scope since it was only a proof of concept.

Furthermore, the whole process from incident data being submitted to the CARPA platform into the visualisation dashboard is done by hand. GraphDB has a Workbench REST API that aids

[27] https://www.mining-technology.com/news/collapse-gold-mine-drc-12-miners/

to automate certain steps in the process [28] which streamlines the process.

Since the IPIS research dataset was of great value for this project, the CARPA team could potentially inform the institute about this research and discuss common interests. Shared knowledge between both parties could lead to discussions of the places with many incidents occurring and make a joint effort in raising awareness to these mines which benefits the growth of the project.

Finally, the IPIS dataset contains over 50 attributes per visit. In future work, more attributes could be involved in the KG.

5.2 Challenges & Limitations

Now that the future work is elaborated, there are still some challenges and limitations to the current end results of the artefacts, which are described here.

At this point, there is no overlap with the Wikidata dataset. However, it is beneficial for future work since new incidents could have overlap. The domain expert noted that it is difficult to find the company involved at the mines, since a mine could have two or three different names for the same mining site due to joint ventures (See Appendix A). Furthermore, it is not evident who investors of mining sites are. By knowing who the investors are, producers of end products are found. However, this project is not to was not able to enrich the data with any of these attributes.

The limitations of the project are mainly attributed to the integration of the individual artefacts. However, this is outside the scope of this project and has to be integrated in further iterations.

The competency questions in this research are developed together with the CARPA team. The aim of the questions was to be broad and cover all elements of interest of the graph. However, in coming iterations other questions could arise to which the current structure can not provide an answer. Then, the ontology must potentially be altered, and new triples need to be created without conflicting the current triples.

Another challenge of design science, is that it is hard to estimate, whether a well-argued design decision is indeed the best decision. For instance, some values of the mine attribute are values in the location attribute as well. This means that the mine column consists of the same value, where another mine has this value in the location column. Distinctions must be made in order to prevent confusion. Furthermore, in the dashboard, the choice was made to only visualize the most recent IPIS visits. This eliminates insights on other earlier visits, which could be interesting when for instance another armed group was present at an earlier visit.

Next, the choices for the data model, are based upon the competency questions provided by the domain experts of the CARPA project. However, there is another stakeholder, local NGOs in the DRC. In this project, it was not possible to interact with local NGOs, so the competency questions are developed partly upon estimations of the needs of the NGOs. Still, the domain experts have collaborated with the NGOs in the previous years with in the case of one expert who has over 15 years of experience in the field. Therefore, the estimation may be assumed as plausible.

Finally, the KG and dashboard can not be openly published due to privacy concerns and sensitive data.

References

[1] Bilal Abu-Salih. “Domain-specific knowledge graphs: A survey”. In: *Journal of Network and Computer Applications* 185 (2021), p. 103076.

[2] Kamal Azzaoui et al. “Scientific competency questions as the basis for semantically enriched open pharmacological space development”. In: *Drug Discovery Today* 18.17-18 (2013), pp. 843–852.

[3] Richard L Baskerville. “Investigating information systems with action research”. In: *Communications of the association for information systems* 2.1 (1999), p. 19.

[4] Camila Bezerra, Fred Freitas, and Filipe Santana. “Evaluating ontologies with competency questions”. In: *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*. Vol. 3. IEEE. 2013, pp. 284–285.

[5] Dan Brickley, Ramanathan V Guha, and Andrew Layman. “Resource description framework (RDF) schema specification”. In: (1999).

[28] https://graphdb.ontotext.com/documentation/standard/using-the-workbench-rest-api.html

[6] Robert Masua Bwana et al. "Developing a Crowdsourcing Application for Responsible Production in Africa". In: *12th ACM Conference on Web Science Companion*. 2020, pp. 48–53.
[7] Haihua Chen et al. "A Practical Framework for Evaluating the Quality of Knowledge Graph". In: *Knowledge Graph and Semantic Computing: Knowledge Computing and Language Understanding*. Ed. by Xiaoyan Zhu et al. Singapore: Springer Singapore, 2019, pp. 111–122. ISBN: 978-981-15-1956-7.
[8] Dalila Cisco Collatto et al. "Is action design research indeed necessary? Analysis and synergies between action research and design science research". In: *Systemic Practice and Action Research* 31.3 (2018), pp. 239–267.
[9] UN Desa et al. "Transforming our world: The 2030 agenda for sustainable development". In: (2016).
[10] Dieter Fensel et al. "Introduction: What Is a Knowledge Graph?" In: *Knowledge Graphs: Methodology, Tools and Selected Use Cases*. Cham: Springer International Publishing, 2020, pp. 1–10. ISBN: 978-3-030-37439-6. DOI: 10.1007/978-3-030-37439-6_1. URL: https://doi.org/10.1007/978-3-030-37439-6_1.
[11] R. V. Guha, Dan Brickley, and Steve Macbeth. "Schema.Org: Evolution of Structured Data on the Web". In: *Commun. ACM* 59.2 (Jan. 2016), pp. 44–51. ISSN: 0001-0782. DOI: 10.1145/2844544. URL: https://doi-org.vu-nl.idm.oclc.org/10.1145/2844544.
[12] Alan R Hevner et al. "Design science in information systems research". In: *MIS quarterly* (2004), pp. 75–105.
[13] Paul Johannesson and Erik Perjons. *An introduction to design science*. Springer, 2014.
[14] Mayank Kejriwal. *Domain-specific knowledge graph construction*. Springer, 2019.
[15] David Nadeau and Satoshi Sekine. "A survey of named entity recognition and classification". In: *Lingvisticae Investigationes* 30.1 (2007), pp. 3–26.
[16] Srinivas Nidhra and Jagruthi Dondeti. "Black box and white box testing techniques-a literature review". In: *International Journal of Embedded Systems and Applications (IJESA)* 2.2 (2012), pp. 29–50.
[17] Natalya F Noy, Deborah L McGuinness, et al. *Ontology development 101: A guide to creating your first ontology*. 2001.
[18] Heiko Paulheim. "Knowledge graph refinement: A survey of approaches and evaluation methods". In: *Semantic web* 8.3 (2017), pp. 489–508.
[19] Michael P Peterson. "Evaluating mapping APIs". In: *Modern Trends in Cartography*. Springer, 2015, pp. 183–197.
[20] Stijn Schouten et al. "The wind in our sails: developing a reusable and maintainable Dutch maritime history knowledge graph". In: *Proceedings of the 11th on Knowledge Capture Conference*. 2021, pp. 97–104.
[21] Ian Taylor. "Dependency redux: Why Africa is not rising". In: *Review of African Political Economy* 43.147 (2016), pp. 8–25.
[22] Mike Uschold and Michael Gruninger. "Ontologies: Principles, methods and applications". In: *The knowledge engineering review* 11.2 (1996), pp. 93–136.
[23] Andra Waagmeester et al. "Science Forum: Wikidata as a knowledge graph for the life sciences". In: *Elife* 9 (2020), e52614.
[24] Dawid Wiśniewski et al. "Analysis of ontology competency questions and their formalizations in SPARQL-OWL". In: *Journal of Web Semantics* 59 (2019), p. 100534.

A Interview Francois Lefant 23-02-2022

What is your function in the CARPA project?
My function in the CARPA project is looking at how could we create links with existing companies from the information provided by the CARPA users. By connecting dots of information and intel I sometimes received, I developed a database with the names of companies in the Democratic Republic of Congo (DRC) with their locations so that I could match basic background information.

What information is currently provided to you from the CARPA application (the raw data)?
The information I receive from the CARPA application is anything related to incidents or initia-

tives in mines. It's a very basic form. We've decided to have a qualitative block to ensure that people write it down without having to select various boxes. That choice is made because the characterization of a selection of items was difficult for the users and can be tedious. During testing phases in Mali and Rwanda a couple of years ago, we realized that the people who we want to fill in the forms, we're more inclined to write like Facebook or WhatsApp. This kind of writing is familiar to writing little text using their own words instead of choosing from a particular categorization that may not be accurate, and which may be time-consuming for our users. It means more work for us processing the data, but the effect is that the workers are more willing to fill in the form.

We started to keep it as open and as wide as possible. Even though we want many incidents linked to child labour, human rights violations, the presence of armed groups, and corruption. However, we can only link them to a particular company to some extent, so that we can start lobby work. We try to train the workers, and we tell them to be as specific as possible. Unfortunately, there's a difference in terms of values and way of working. What I find specific and interesting differs from what the workers find interesting and specific.

What adjustments do you make to the raw data?
I added the mining site, location, country, local, and international company name. The CARPA application provides me with the case name, type, and the column incident/initiative.

What information is currently missing and could be enriched in your opinion?
The problem is that in only a few cases you can say very clearly which company you are dealing with. In the mining industry, one must know two or three different names for the same site, since there are many joint ventures. That is why a knowledge graph containing, for instance, a different name, can be linked to the same mining site when used in CARPA submissions. Something I would like to see is an additional column including investors so that pressure can be put on them as well. The difficult part is maintaining the information. The path from the mining site to the local company, to the joint venture, to the actual company and finally to the investors, requires extensive research. Furthermore, such information changes on a half-year or year basis.

Further enrichment to the sheet could be the addition of raw materials mined on a certain site. In a certain village, there can be various raw materials mined. Therefore, when a CARPA user fills in an incident and provides a village name, the material gained, and the site is still not clear.

What case-specific knowledge do you use to make certain decisions?
I've been to the DRC many times over the past 20 years, so I know some companies and some things happening. So, it's either my knowledge or just reading the documents in depth when these are attached. Sometimes a basic search on the Internet can help me figure out what company is where. All these methods help me fill in the gaps. I am always looking from the perspective of an outsider. The linkage between local and international is the added value of CARPA, but simultaneously the tricky part.

B Outcomes dashboard test

	User 1	User 2	User 3	User 4	User 5
1.1 How many incidents are related to a specific type of mineral?	Found the 6 mines	Found the 6 gold mines by filtering incident_type and mineral type	Found the 6 gold mines	Found the gold mines	Found the mines, took a long time to understand that there was a table below
1.2 What violation type is most common with extracting a particular type of mineral	Physical-Abuse, found by counting markers on the map	Physical-Abuse	Filtered on gold mines and looked at the table. Next, counted the rows of the most frequent	Physical-Abuse	Physical-Abuse
2.1 How many incidents have been raised relating to this company over a period of time?	Possible but no clear overview	Possible but no clear overview	Possible but no clear overview	Possible but no clear overview	Possible but no clear overview
2.2 How many incidents have been raised related to the extraction of a specific mineral over a period of time?	Possible but no clear overview	Possible but no clear overview	Possible but no clear overview	Possible but no clear overview	Possible but no clear overview
2.3 How many incidents have occurred over a defined period of time?	Could be done via table but was not in chronological order, therefore difficult	Could be done via table but was not in chronological order, therefore difficult	Possible but no clear overview	Possible but no clear overview	Done
2.4 what is the comparison of incident numbers between one period and another	Possible but no clear overview	Possible but no clear overview	Possible but no clear overview	Possible but no clear overview	Possible but no clear overview
3.1 How many incidents are related to a mineral type within a region?	-	-	-	-	-
3.2 How many incidents in a region are of a particular violation type?	-	-	-	-	-
3.3 How are incidents spread between regions or within regions?	-	-	-	-	-
3.4 How many incidents were reported in the region of Nord-Kivu in 2020? & Query mines where incidents occurred, including the incident date and the region. Next, filter by Nord-Kivu and order by date.	-	-	-	-	-
4.1 How many incidents are related to a particular company? & Query mines where incidents occurred, including the mine operator	Possible but no clear overview	Possible but no clear overview	Possible but no clear overview	Possible but no clear overview	Possible but no clear overview
4.2 Which violation type is most often reported regarding a particular company? & Query mines where incidents occurred, including the mine operator and the incident type	Possible but no clear overview	Possible but no clear overview	Possible but no clear overview	Possible but no clear overview	Possible but no clear overview
5.1 What armed groups are present at mines where incidents occurred? & Query mines where incidents occurred and was visited, including armed groups	FARDC	Had some struggle to find column, did it eventually	FARDC	FARDC	FARDC
6.1 Is there child labour at mines where incidents occurred? & Query mines where incidents occurred and was visited, including child labour	-	-	-	-	-
6.2 How many people work at mines where incidents occurred? & Query mines where incidents occurred and was visited, including women	-	-	-	-	-
General comments	Long descriptions extend the assigned text box	Mineral and Had_visit do not filter back to all when clicked.	Not clearly visible when mines are at the same location	Tried to filter columns and expected it to work but did not do this.	Did not realise data changed on the right, expected a pop-up from the marker it self
	Clusters do not reload when filtering	Chronological order of date created in table desired	Region column would be interesting as well as a region filter column		Possible legenda to explain the colours
			Sort data in the table		

Fig. 10: Dashboard black box testing results