Knowledge Engineering and Ontology for Crime Investigation

Wilmuth Müller[1](✉), Dirk Mühlenberg[1], Dirk Pallmer[1], Uwe Zeltmann[1], Christian Ellmauer[1], and Konstantinos Demestichas[2]

[1] Fraunhofer IOSB, Fraunhoferstr. 1, 76131 Karlsruhe, Germany
wilmuth.mueller@iosb.fraunhofer.de
[2] Institute of Communication and Computer Systems, Iroon Polytechneiou 9, 15773 Zografou, Greece
cdemest@cn.ntua.gr

Abstract. Building upon the possibilities of technologies like ontology engineering, knowledge representational models, and semantic reasoning, our work presented in this paper, which has been performed within the collaborative research project PREVISION (Prediction and Visual Intelligence for Security Information), co-funded by the European Commission within Horizon 2020 programme, is going to support Law Enforcement Agencies (LEAs) in their critical need to exploit all available resources, and handling the large amount of diversified media modalities to effectively carry out criminal investigation.

A series of tools have been developed within PREVISION which provide LEAs with the capabilities of analyzing and exploiting multiple massive data streams coming from social networks, the open web, the Darknet, traffic and financial data sources, etc. and to semantically integrate these into dynamic knowledge graphs that capture the structure, interrelations and trends of terrorist groups and individuals and Organized Crime Groups (OCG).

The paper at hand focuses on the developed ontology, the tool for Semantic Reasoning and the knowledge base and knowledge visualization.

Keywords: Ontology · Knowledge base · Semantic reasoning · Knowledge visualisation

1 Introduction

Organised Crime Groups (OCGs) quickly adopt and integrate new technologies into their 'modi operandi' or build brand-new business models around them (such as CaaS) [1]. More than 5,000 OCGs operating on an international level are currently under investigation in the EU, whereas document fraud, money laundering and the online trade in illicit goods and services are recognised as the engines of organised crime. Notably, goods and services offered on the Darknet are available to anyone, be it an individual user, an OCG or terrorist organization [2].

Published by Springer Nature Switzerland AG 2022
I. Maglogiannis et al. (Eds.): AIAI 2022, IFIP AICT 646, pp. 483–494, 2022.
https://doi.org/10.1007/978-3-031-08333-4_39

This calls for new tools that allow Law Enforcement Agencies (LEAs) to understand the structure, complexity, dynamics and interrelations within and across OCGs or terrorist organisations. It is important to provide LEAs advanced Big Data capabilities and appropriate Information and Communication Technology (ICT) tools that analyse social networks, utilising advanced linguistic models and semantic technologies.

The EU funded project PREVISION developed a series of tools which provide LEAs with the capabilities of analysing and exploiting multiple massive data streams. These are coming from social networks, the open web, the Darknet, traffic and financial data sources, etc. These tools semantically integrate the acquired data into dynamic knowledge graphs that capture the structure, interrelations and trends of terrorist groups and individuals and OCGs. The tools have been integrated into and interconnected in a platform, providing LEAs a common access to them. Figure 1 presents an overview of the architecture of the developed platform. The platform has been described in [3].

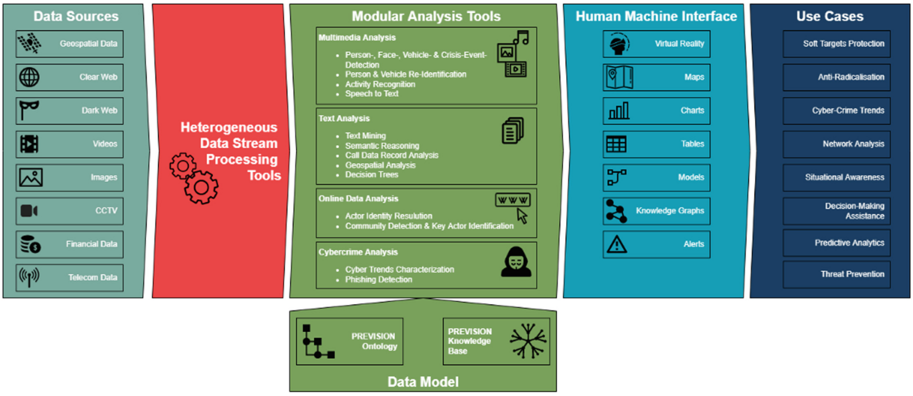


Fig. 1. Overview of the PREVISION platform architecture

This paper focuses on the developed ontology, the semantic reasoning tool, the knowledge base and knowledge visualization tools.

The rest of the paper is structured as follows: Sect. 2 describes related work, especially other EU funded projects addressing the topic of technologies improving the knowledge of investigators when fighting crime and terrorism. Section 3 introduces the PREVISION ontology, which is the core of the PREVISION platform. Section 4 describes how the knowledge base has been implemented. In Sect. 5 semantic reasoning tools, which operate on the knowledge base and produce new entries in it are provided. Section 6 presents how the knowledge is visualized. The paper concludes with a conclusion and the acknowledgment.

2 Related Work

In order to foster the fight against organized crime and terrorism, the European Union funded a series of research projects to develop tools needed by LEAs. These projects are similar to PREVISION, each of it focusing on specific aspects regarding the investigation of criminal and terrorist activities.

The ANITA (Advanced tools for fighting online illegal trafficking) project [4] focused on the design and development of a knowledge-based user-centered investigation system for analyzing heterogeneous (text, audio, video, image) online (surface web, deep web, DarkNet) and offline content for fighting financing terrorism, illegal trafficking of drugs, counterfeit medicines, and firearms.

The COPKIT project [5] has developed data-driven policing technologies to support Law Enforcement Agencies in analysing, investigating, mitigating and preventing the use of new information and communication technologies by organized crime and terrorist groups. It developed a toolkit for knowledge production and exploitation in investigative and strategic analysis work to support the Early Warning /Early Action paradigm for both strategic and operational levels.

The AIDA (Artificial Intelligence and Advanced Data Analytics for Law Enforcement Agencies) project [6] is developing a Big Data Analysis and Analytics framework equipped with a complete set of automated data mining and analytics solutions to deal with standardised investigative workflows, extensive content acquisition, information extraction and fusion, knowledge management and enrichment through applications of Big Data processing, Machine Learning, AI and predictive and visual analytics. It is focusing on cybercrime and terrorism by approaching specific issues and challenges related to LEAs' investigation using machine learning and artificial intelligence methods.

The ASGARD (Analysis System for Gathered Raw Data) project [7] developed a best-of-class tool set for the extraction, fusion, exchange and analysis of Big Data, including cyber-offense data for forensic investigation.

The INSPECTr (Intelligence Network & Secure Platform for Evidence Correlation and Transfer) project [8] develops a shared intelligent platform and a process for gathering, analysing, prioritizing, and presenting key data to help in the prediction, detection and management of crime in support of multiple agencies at local, national and international level. Using both structured and unstructured data as input the developed platform facilitates the ingestion and homogenisation of this data with increased levels of automation, allowing for interoperability between multiple data formats.

The TENSOR (Retrieval and Analysis of Heterogeneous Online Content for Terrorist Activity Recognition) project [9] developed a unified semantic infrastructure for information fusion of terrorism-related content and threat detection on the Web. The TENSOR framework consists of an ontology and an adaptable semantic reasoning mechanism.

In literature there are several attempts to model terrorism-related concepts as ontologies, with the work by Mannes and Golbeck being one of the first attempts [10, 11]. In their work the authors present an ontology for representing terrorist activity, mostly focusing on the description of sequences of events and the representation of social networks underpinning terrorist organizations.

3 Ontology

One of the first steps in the development of the PREVISION toolset has been the design of an ontology capturing the relevant concepts used in criminal investigations.

The model has been defined using the semantic web technology Resource Description Framework (RDF) for describing an ontology. Ontologies are a formal way to describe taxonomies and classification networks, essentially defining the structure of knowledge for various domains. The World Wide Web Consortium (W3C) defines the Web Ontology Language (OWL) as a knowledge representation language for authoring ontologies.

The PREVISION ontology is based on the so-called intelligence pentagram (Fig. 2) that is widely used in the field of intelligence analysis [12]: The pentagram connects the following main concepts:

- Event: A description of an incident or occurrence of some significance that happens during a defined time period. Examples of important PREVISION-specific event types are special crime types, actions in the preparation or execution of a crime, watching a crime by witnesses/testimonies and police counter-crime measures.
- Equipment: Any item of material used to equip a person, organization or place to fulfil its role.
- Organization: An organizational entity or grouping which has a common purpose and which may have a recognizable hierarchical structure.
- Place: Represents all spatial areas, which may be relevant in the context of a crime. A place may be a natural or a man-made feature, an area or a geospatial reference point.
- Person: A description of the physical characteristics and of the private and professional attributes of an individual. This will consist of, amongst other matters, details of the identification, relationships to other persons and digital identities and individual behavior patterns of the person.

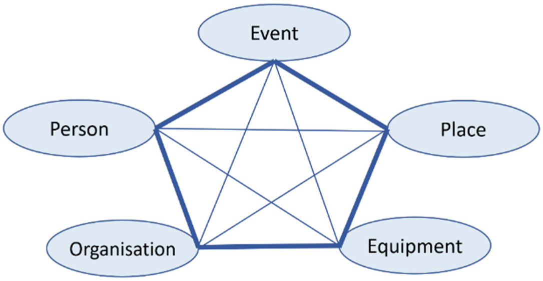


Fig. 2. Main ontology concepts

In order to support information integration and cyber situational awareness in cybersecurity systems, the PREVISION ontology has been enhanced by integrating the Unified Cybersecurity Ontology (UCO). The ontology incorporates and integrates heterogeneous data and knowledge schemas from different cybersecurity systems and most commonly used cybersecurity standards for information sharing and exchange [13].

Significant effort has also been placed on identifying complementary underlying concepts via a data-driven process, by further analyzing the knowledge generated by PREVISION data sources. One of these data sources have been datasets acquired by the Dark Web crawler (see Fig. 1, left part). The analysis of these datasets revealed new concepts related to online marketplace advertisements, user profiles in forums, marketplaces, forum posts and digital currency addresses (e.g. Bitcoin addresses). The concepts have been integrated in the ontology, thus enlarging its coverage of criminal acts.

The PREVISION ontology is compatible with the Universal Message Format (UMF), a standard or agreement on what the structure of the most important law enforcement concepts when they are exchanged across borders should be. UMF is a set of concepts (building blocks) to construct standard data exchanges for interconnecting dispersed law enforcement systems [14].

Based on the PREVISON ontology, a message format for the information exchange between analysis tools has been developed. Messages are formulated in the widely used JSON format, where keys and values of JSON objects are governed by the PREVISION ontology. This approach opens up the possibility to standardize the interfaces for message exchange within the platform and even the platform and external systems in order to establish a modularized, open architecture.

4 Knowledge Base

The output of several tools for data mining, data stream processing, and information extraction is fused in a common knowledge graph. It is represented as a set of triples of the form “subject” – “predicate” – “object” according to the RDF standard and can be queried with the SPARQL Protocol and RDF Query Language (SPARQL). Besides this kind of information which we call “semantic data”, also other kinds of data like text, image, and video files are produced and stored.

As an implementation of the semantic data base, the Apache Jena Fuseki server has been chosen, while binary data is stored in an Apache Hadoop Distributed File System (HDFS) and MongoDB is used as a document-oriented data base. Furthermore, an API encapsulating the Fuseki server has been developed which serves several purposes:

Data consistency: Write operations on the knowledge base are only accepted if the inserted data is consistent with the PREVISION ontology. In particular, the class and property hierarchy as well as domain and range specifications are respected.

Data provenance: The RDF format allows the partition of a knowledge graph into named subgraphs. In the PREVISION knowledge base, the name of a subgraph is an OWL individual itself, which is assigned information to by a set of triples about the system component or user which has inserted the triples in the subgraph. In this way transparency about data provenance is assured. In theory, it would also be possible to encode the point in time in this way, when a triple has been created in the knowledge base. However, in PREVISION this option has not been used.

REST API: The PREVISION knowledge base provides a Representational State Transfer (REST) API which is accessible over the HTTP as well as Advanced Message Queuing Protocol (AMQP) protocols. As an implementation for the latter, PREVISION makes use of the message broker RabbitMQ[1].

[1] https://www.rabbitmq.com/.

Several functions are included in the API:

SPARQL Read Access: For read access, the knowledge base exposes a SPARQL 1.1 Protocol compliant SPARQL endpoint.

Nested JSON Object Interpretation: Data insertions can be made in the form of nested JSON structures, which the API internally translates into sets of RDF triples. Keys and values of objects within these structures are checked to be compliant with the PREVISION ontology.

As an example, the JSON object

```
{ "a": "EmailAccount",
  "isAccountOf": {
           "a": "Person",
           "hasPersonSurname": "Smith"}
}
```

resolves in the four RDF triples

```
pv:ie_3 a pv:EmailAccount
pv:ie_3 pv:isAccountOf pv:ie_4
pv:ie_4 a pv:Person
pv:ie_4 pv:hasPersonSurname "Smith"
```

where "`pv:`" denotes the PREVISION namespace prefix and `pv:ie_3` and `pv:ie_4` are automatically generated IRIs. To identify several automatically created IRIs, objects can be assigned temporary names to by key/value pairs of the form `"TAG": "temporary name"`. Multiple values for properties can be listed as JSON arrays, and inverse properties can be expressed with the syntax "`inverse(property_name)`".

Graph Node and Property Value Access: Furthermore, the knowledge base API includes a service providing an overview as well as editing facilities of information directly related to a user selected node in the knowledge graph. For a selected entity, all RDF triples in the knowledge graph adjacent to it can be returned. Moreover, in case of the entity being an individual, properties, of which the given individual is an element of the domain or range, together with possibly empty value lists are included in the output. The names of the graphs containing the returned information as well as some technical annotations required by the knowledge base GUI described in Sect. 6.2 are included in the result, too.

In addition, ontology compliant editing options of individuals and their property values, and according write access to create, add, or delete individuals and values of properties is implemented as well.

Case Management: PREVISION separates information belonging to independent crime cases from each other. In this context, the knowledge base provides functions to create, delete and clear case data sets.

5 Semantic Reasoning

PREVISION's semantic reasoning toolset consists of a logic reasoning tool and a probabilistic reasoning tool. The logic Reasoning Tool is based on Semantic Web Rule Language (SWRL). It extracts information from the knowledge base focused on specific SPARQL requests using Fuseki's inbuilt reasoner to apply the ontology-inherent rules resulting from the taxonomy of classes. The results are displayed in tabular or graphical node-network presentation. The various queries include the following aspects:

- Persons and attributes. i.e. vehicle owner/holders, residence, guns/weapons
- Vehicles (route planning)
- Events
- Crisis Event

The probabilistic reasoning tool uses a semantic reasoning technique for extending existing information with new knowledge by adding additional relations between persons, events, places or objects in the knowledgebase. It is based on the so-called Markov Logic Networks (MLN) [15], which enables probabilistic reasoning by combining a probabilistic graphical model with first-order logic.

An MLN represents a first-order knowledge base, i.e. a set of formulas expressed in first-order logic. MLNs have been introduced in 2006 by M. Richardson and P. Domingos, see [15]. Since then they have been an active area of research and were widely applied in different scenarios, e.g. ontology matching, statistical learning and probabilistic inference, see [16–18]. The advantage of probabilistic reasoning is the capability to deal with uncertainties in the knowledge and the rules that the reasoning is applied on.

An MLN is a first-order knowledge base, i.e. a set of formulas $\{F_i \mid i \in I\}$ stated in first-order logic (FOL), where every formula is equipped with a corresponding weight $\omega_i \in \mathbb{R}$. The weight assigned to each formula expresses the degree of believe that the formula is correct. These formulas serve as the base of a procedure yielding a Markov network, which then can be used to assign probabilities to possible states of instances of the underlying ontology.

The set of formulas consists of the evidence retrieved from the knowledge base and the rules that have to be developed in cooperation with law enforcement agencies. The advantage of the probabilistic reasoning is the ability to cope with rules that have uncertainties, meaning they may not hold in every case, but in most cases. So, the confidence in a rule is expressed in a weight factor. As a result, the new evidence inferred has a weight that may be seen as a measure for the probability that it is correct.

The probabilistic reasoning module in PREVISION is based on the open-source implementation of the MLN reasoner Tuffy, developed by Stanford University, which achieves a better scalability than other MLN implementations (see [19]).

In [20] the MLN implementation Tuffy has been integrated into an information fusion component for fusing information acquired by a distributed surveillance system with prior information contained in intelligence databases. Information given in the form of an OWL ontology, such as a taxonomy of defined concepts as well as relations, have proven to be easily convertible into FOL formulas and integrable into an MLN model.

The MAGNETO project [21] developed an MLN based reasoning module for generating new knowledge from witness statements that may contain unreliable information [22].

An adapter has been developed to integrate the MLN reasoner into the PREVISION framework. The adapter connects to the knowledge stored in the Fuseki RDF Triple Store for input and output (see Fig. 3).

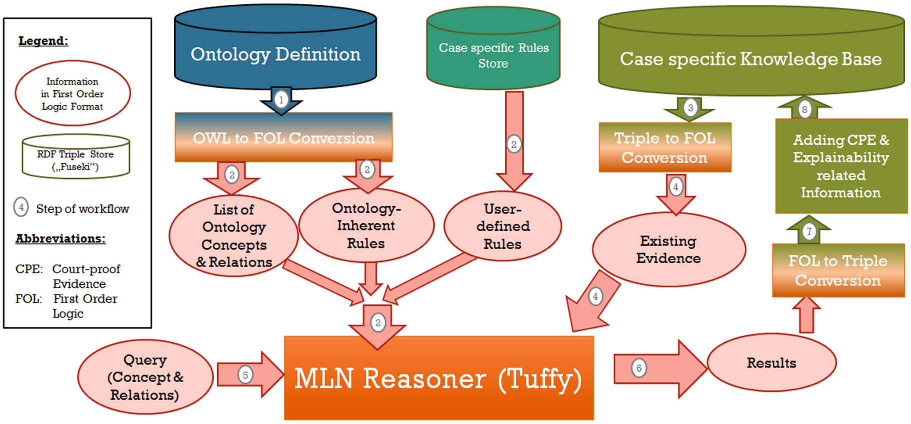

Legend:
Information in First Order Logic Format
RDF Triple Store („Fuseki")
4 Step of workflow
Abbreviations:
CPE: Court-proof Evidence
FOL: First Order Logic
Ontology Definition
Case specific Rules Store
Case specific Knowledge Base
1
OWL to FOL Conversion
2
2
2
3
8
Triple to FOL Conversion
Adding CPE & Explainability related Information
List of Ontology Concepts & Relations
Ontology-Inherent Rules
User-defined Rules
4
Existing Evidence
7
FOL to Triple Conversion
4
2
Query (Concept & Relations)
5
MLN Reasoner (Tuffy)
6
Results

Fig. 3. Workflow of the MLN reasoning in PREVISION

6 Visualisation of Knowledge

Tools in the PREVISION platform like the text mining tool produce results on a high information density level, which corresponds to the information density in the source data. But this data density is not appropriate for an investigator, who is interested in certain indicators related to the case he is in charge of. A central aspect during a criminal investigation is evidence discovery to support some of the hypothesis the investigator has in mind.

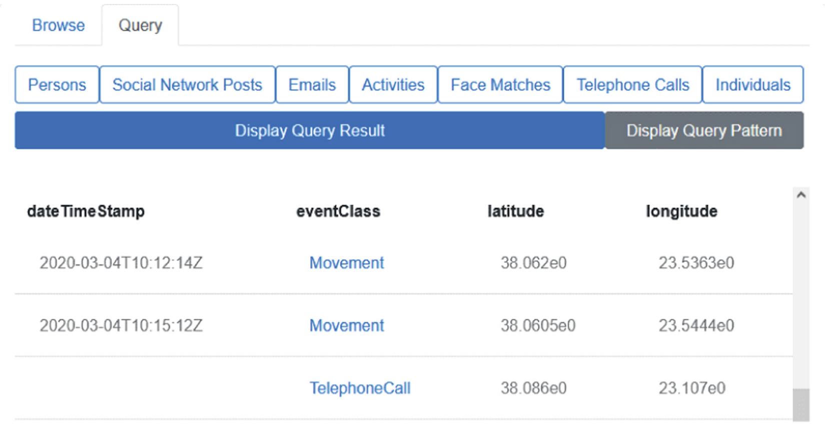
To gain condensed and problem-oriented information, PREVISION has developed a web-based graphical user interface providing several visualization tools to give a comprehensive view on the evidence gathered in the knowledge base by various tools and users. These visualization tools are incorporated in the Knowledge Base Inspector, a web-based graphical user interface (GUI) as part of the overall PREVISION GUI.

6.1 Table View

The Table View of the Knowledge Base Inspector allows the user to run custom SPARQL queries as well as to choose predefined SPARQL query patterns for frequently performed query tasks (see Fig. 4). Also, for each OWL entity stored in the knowledge base, it is possible to list all triples adjacent to it in tabular form.

The display of OWL entities is implemented as web-links in the Table View, which allows the user to browse through the knowledge graph. Also type-specific editing options are provided to create, insert and delete data, while consistency with the ontology is assured.

The Table View further provides the option to display text, image or video files stored in Hadoop which are associated via a dedicated OWL property with the individuals stored in the knowledge base. Also, context menus associated with entities provide the option to run several context specific analysis processes or data visualization tools.



dateTimeStamp	eventClass	latitude	longitude
2020-03-04T10:12:14Z	Movement	38.062e0	23.5363e0
2020-03-04T10:15:12Z	Movement	38.0605e0	23.5444e0
	TelephoneCall	38.086e0	23.107e0

Fig. 4. Results of a SPARQL-query for geo- and time-referenced data. By clicking on the events Movement or TelephoneCall, these could be further investigated.

6.2 Graph View

The Knowledge Graph Visualizer is a data visualization tool incorporated into the Knowledge Base Inspector offering a graphical view on the linked data of the knowledge base. The tool is designed as a REST[2]-API. It combines the power of SPARQL queries via the Knowledge Base API with state-of-the-art software packages for working with RDF data like RDFLib [23] and Networkx [24], providing algorithms for graph-analysis and manipulation.

Based on a selection of classes and individuals or by providing a list of search terms, the RDF-triples (subject, predicate and object) of their instances are queried. The subjects and objects result in nodes of a directed graph, connected via predicates representing the edges. Subjects and objects which are instances of a certain class are depicted in blue whereas literal objects are depicted in yellow. To further simplify the view on the linked data, technical details which are not relevant to the end user are removed from the view.

A more comprehensive view on the requested data provides a so-called ego-graph where the result entity is used as a center node, surrounded by a number of nodes containing associated data. These depend on a radius to be specified, which defines the distance in nodes to be displayed around the center-node.

The following excerpt depicts the nodes and edges from an ego-graph with a radius of three around “TelephoneCall”, “Person”, and “Movement” instances in the knowledge

[2] Representational State Transfer.

base (Fig. 5). The instances are the same as the one displayed via the Table View and Map-based View.

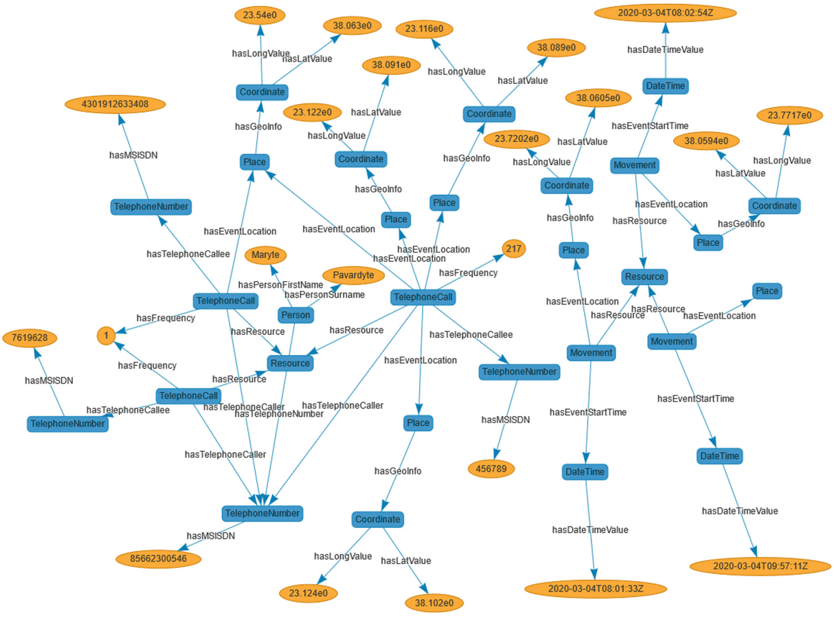


Fig. 5. Graph-based view on the knowledge base data.

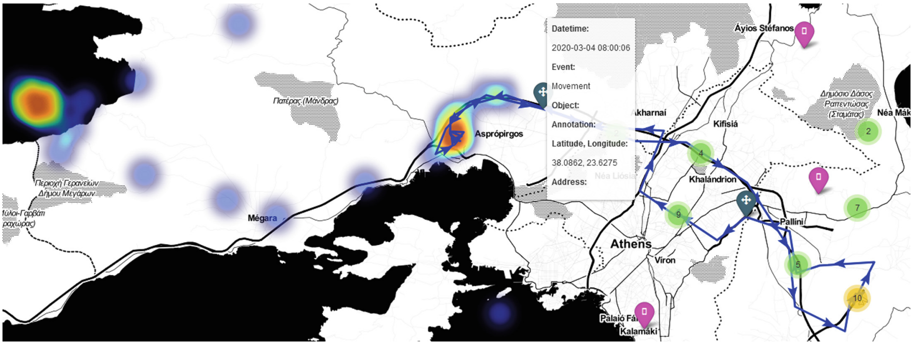
6.3 Map-Based View

Events that have a geo and an optional time reference may be displayed in an interactive map that the user can explore. Different visualization types will help an investigator to get an overview of the event-locations which can be either displayed as simple markers on a map or as heatmaps in order to get an impression of the data distribution. Depending on the zoom-level and density of markers, they are clustered into marker-clusters representing a number of events in a certain area.

With an additional time-component, events may be sorted in chronological order and connected with each other with arrows to form trajectories. Connection maps show the locations of two simultaneous events connected by a line, for example the locations of two persons involved in a telephone call.

Finally, a timeseries analysis provides the ability to analyze a sequence of events collected over a specific recording duration where the events are then aggregated over a set period of time (e.g. hourly, daily, etc.) in order to present a specific heatmap or marker cluster for each period of the whole recording duration as an animation. An example may be to analyze the development of monthly burglaries over a recording duration of ten years.

The following illustration shows a heatmap and marker cluster of the knowledge base containing a "Person" instance who produced some geo-referenced "TelephoneCall" and additionally timestamped "Movement" events resulting in a trajectory (Fig. 6).



Datetime:
2020-03-04 08:00:06
Event:
Movement
Object:
Annotation:
Latitude, Longitude:
38.0862, 23.6275
Address:

Fig. 6. Map-based view on the knowledge base data.

7 Conclusion

In this paper, a semantic framework for knowledge management is presented to support LEAs by improving their situation awareness in criminal investigation.

Various analysis tools have been developed to extract information from heterogenous data. The output of the information extraction components is used to populate a knowledge base structured by an ontology, which has been developed specifically for crime investigation purposes.

The consequent usage of the PREVISION ontology as the basis for an exchange format of messages between platform components, which can themselves be interpreted and persisted as parts of a knowledge graph, turned out to be a promising approach for a platform architecture in the domain of criminal investigation.

The visualization of the knowledge base empowers the investigators to gain an enhanced overview and situation awareness of the case under investigation. The paper presented three different views on the same data set in the knowledge-base.

The PREVISION ontology has been asked by and offered to other research projects funded by the European Union in the domain of fight against organized crime and terrorism.

Acknowledgments. This work has been performed under the H2020 833115 project PREVISION, which has received funding from the European Union's Horizon 2020 Programme. This paper reflects only the authors' view, and the European Commission is not liable to any use that may be made of the information contained therein.

References

1. Europol. Serious and Organised Crime Threat Assessment (SOCTA) (2017)
2. Europol. Internet Organised Crime Threat Assessment (IOCTA) (2017)
3. Demestichas, K., et al.: Prediction and visual intelligence platform for detection of irregularities and abnormal behaviour. In: Detection Machine Learning for Trend and Weak Signal Detection in Social Networks and Social Media, vol. 2606, no. paper 4, pp. 25–30. CEUR (2020)

4. ANITA project. https://www.anita-project.eu/
5. COPKIT project. https://copkit.eu/
6. AIDA project. https://www.project-aida.eu
7. ASGARD project. https://www.asgard-project.eu
8. INSPECTr project. https://inspectr-project.eu
9. TENSOR project. https://tensor-project.eu
10. Mannes, A., Golbeck, J.: Building a terrorism ontology. In: ISWC Workshop on Ontology Patterns for the Semantic Web, vol. 36 (2005)
11. Mannes, A., Golbeck, J.: Ontology building: a terrorism specialist's perspective. In: Aerospace Conference, 2007 IEEE, pp. 1–5. IEEE (2007)
12. Dragos, V.: Developing a core ontology to improve military intelligence analysis. Int. J. Knowl.-Based Intell. Eng. Syst. **17**, 29–36 (2013). https://doi.org/10.3233/KES-130253
13. Syed, Z., Padia, A., Finin, T., Mathews, L., Joshi, A.: UCO: a unified cybersecurity ontology. In: AAAI Workshop on Artificial Intelligence for Cyber Security 2016. AAAI Press (2016). https://doi.org/10.13016/M2862BG1V
14. Europol. Universal Message Format: faster, cheaper, better. Publications Office of the European Union. https://op.europa.eu/en/publication-detail/-/publication/3b2cc49f-72bb-419f-8742-eb21cd15e35c
15. Richardson, M., Domingos, P.: Markov logic networks. Mach. Learn. **62**(1–2), 107–136 (2006). https://homes.cs.washington.edu/~pedrod/mln.pdf
16. Niepert, M., Meilicke, C., Stuckenschmidt, H.: A probabilistic-logical framework for ontology matching. In: AAAI (2010)
17. Chen, H., et al.: Scaling up Markov logic probabilistic inference for social graphs. IEEE Trans. Knowl. Data Eng. **29**(2), 433–445 (2017)
18. Wittek, P., Gogolin, C.: Quantum enhanced inference in Markov logic networks. Sci. Rep. **7**, 45672 (2017)
19. Niu, et al.: Tuffy: scaling up statistical inference in Markov logic networks using an RDBMS. Proc. VLDB Endow. **4**(6), 373–384 (2011)
20. Kuwertz, A., Mühlenberg, D., Sander, J., Müller, W.: Applying knowledge-based reasoning for information fusion in intelligence, surveillance, and reconnaissance. In: Lee, S., Ko, H., Oh, S. (eds.) MFI 2017. LNEE, vol. 501, pp. 119–139. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-90509-9_7
21. MAGNETO homepage. http://www.magneto-h2020.eu/. Accessed 09 Feb 2022
22. Müller, W., et al.: Reasoning with small data samples for organised crime. In: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II, (S. 21) (2020). https://doi.org/10.1117/12.2557543
23. RDFLib software package. https://rdflib.readthedocs.io/. Accessed 11 Feb 2022
24. NetworkX software package. https://networkx.org/. Accessed 11 Feb 2022