CEM: an Ontology for Crime Events in Newspaper Articles

Federica Rollo
Enzo Ferrari Engineering Department
University of Modena and Reggio Emilia
Modena, Italy
federica.rollo@unimore.it

Laura Po
Enzo Ferrari Engineering Department
University of Modena and Reggio Emilia
Modena, Italy
laura.po@unimore.it

Alessandro Castellucci
Enzo Ferrari Engineering Department
University of Modena and Reggio Emilia
Modena, Italy
228058@studenti.unimore.it

ABSTRACT

The adoption of semantic technologies for the representation of crime events can help law enforcement agencies (LEAs) in crime prevention and investigation. Moreover, online newspapers and social networks are valuable sources for crime intelligence gathering. In this paper, we propose a new lightweight ontology to model crime events as they are usually described in online news articles. The Crime Event Model (CEM) can integrate specific data about crimes, i.e., where and when they occurred, who is involved (author, victim, and other subjects involved), which is the reason for the occurrence, and details about the source of information (e.g., the news article). Extracting structured data from multiple online sources and interconnecting them in a Knowledge Graph using CEM allow events relationships extraction, patterns and trends identification, and event recommendation.

The CEM ontology is available at https://w3id.org/CEMontology.

CCS CONCEPTS

• **Information systems → Semantic web description languages**; **Ontologies**.

KEYWORDS

lightweight ontology, crime analysis, newspaper

ACM Reference Format:
Federica Rollo, Laura Po, and Alessandro Castellucci. 2023. CEM: an Ontology for Crime Events in Newspaper Articles. In *The 38th ACM/SIGAPP Symposium on Applied Computing (SAC '23), March 27 – March 31, 2023, Tallinn, Estonia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3555776.3577862

1 INTRODUCTION

Leveraging event data to derive insights is crucial to make effective decisions in several contexts, e.g., advertising, resource planning, price strategies, and, also, crime prevention. To facilitate the semantic analytics of information regarding events on the Web, a comprehensive representation of events, entities, spacial, causal and temporal relations provided by a Knowledge Graph (KG) or an ontology is needed. Existing KGs, such as Wikidata, DBpedia and YAGO, focus mostly on entity-centric information, moreover, the representation is limited to fundamental and historical events. When it comes to express daily events described in news, or social media content, event-centric KGs are able to capture the dynamic aspects of events and allow representing the relationships between them, building temporal and causal chains.

In the domain of Crime Analysis, as defined by the International Association of Crime Analysts (IACA),[1] qualitative and quantitative techniques are employed to analyze data useful to LEAs, including the analysis of actual crimes, criminals, victims, problems related to the quality of life, socio-demographic aspects and other factors that can influence the crime occurrences in a city. The result of these analyses is aimed at preventing crime, supporting criminal investigations and evaluating the actions of LEAs. Usually to represent crime information, ontologies based on the national judicial system or connected to the crime investigation process are exploited. This is the case of the Italian Crime Ontology [2]. Even if they are invaluable to a deep analysis of the crime process to support examiners in forensic investigations, they are complex and detailed formalization of crime events, that are not so detailed when it comes to describe them in a news or on a post in social media.

In the literature, for the best of our knowledge, there is no ontology for the representation of crimes extracted from online newspapers. We identified some relevant works in the domain of event representation: the multilingual event-centric temporal KG, named EventKG [6], for the fusion of events and temporal relations from large-scale entity-centric KGs, based on the Simple Event Model (SEM) [13], the Open Event Knowledge Graph [7], which extends EventKG to integrate heterogeneous data and exploits the data model Schema.org to link each event to the source describing that event, and the Entity-Event Knowledge Graph [1] developed for business scopes where entities and events are organized hierarchically and concepts are linked to elements in ontologies, taxonomies and thesauri. However, the existing ontologies are not suitable to give a comprehensive representation of a crime event due to the lack of specific classes and properties.

In this paper, a new lightweight ontology, named *Crime Event Model* (CEM), is defined for the representation of crime events. The scope of the model is the management of information related to crime events extracted from newspaper articles, trying to preserve semantic interconnections with existing KGs and ontologies following the FAIR principles [9, 14]. The model is independent of the legal representation of crime (that differs from country to country) and is based on the fundamental 5W1H journalistic questions which are related to: object(s) (e.g., `What` has been stolen?, `What` object was used to commit the crime?), participant(s) (i.e., `Who` was involved?, `Who` was the victim?, `Who` was the author of the crime?),

ACM SAC '23, March 27 – March 31, 2023, Tallinn, Estonia

https://doi.org/10.1145/3555776.3577862

[1] https://iaca.net/

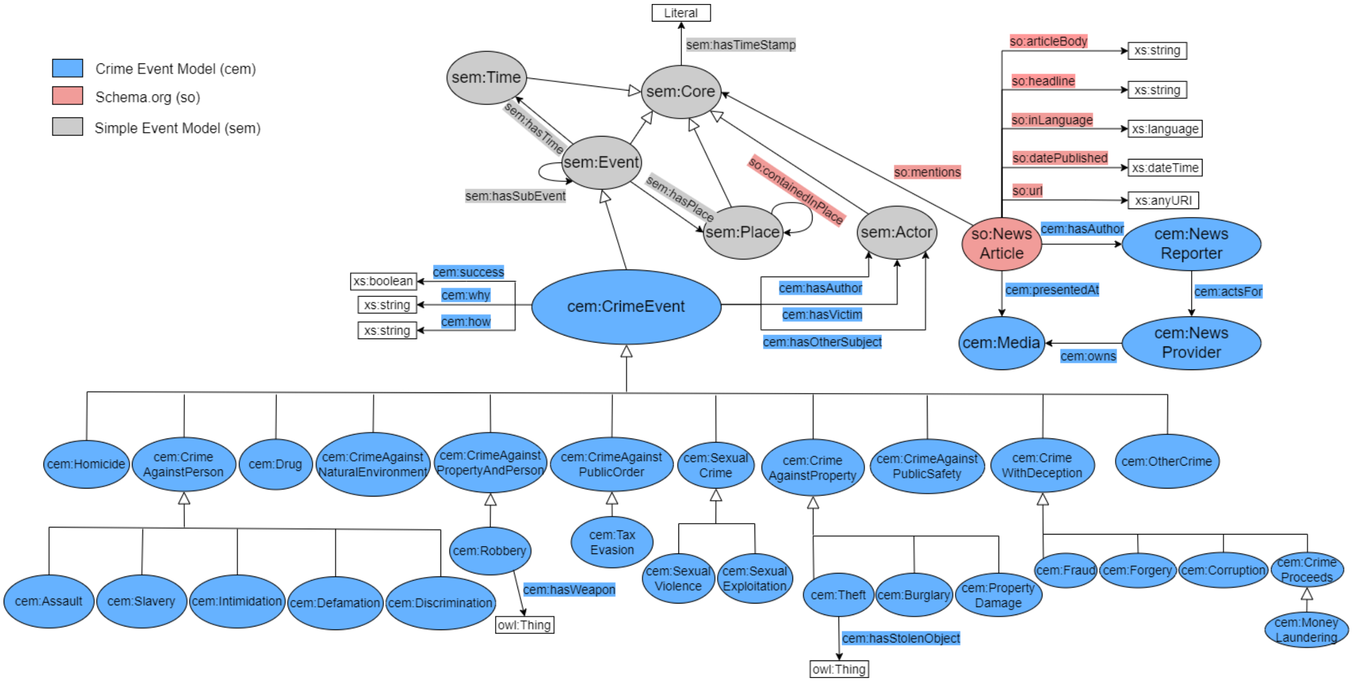

Figure 1: The Crime Event Model (CEM).

place (i.e., Where did the crime take place?), time (i.e., When did it take place?), cause (i.e., Why did it happen?), procedure (i.e., How did it happen?). The representation of crimes in such a model could support decision-making systems integrating data from multiple data sources and providing a semantic infrastructure to them.

2 CRIME EVENT MODEL

The scope of CEM is to enrich semantically information contained in news articles and represent those data in such a model that can be useful to make analysis on crime events. We propose a new ontology that extends existing KGs and ontologies to provide details on crime events. The reuse of existing models allows more interoperability and interconnection. In particular, *SEM* and *Schema.org* have been taken into account with the possibility to exploit connections to DBpedia, Wikidata and other well-known KGs.

2.1 Requirements

Before the design of the ontology, we selected some requirements: 1) modeling different types of crimes, 2) keeping a reference to news articles reporting the crime events, 3) modeling the subjects of the crimes according to their roles, 4) modeling the objects of the crimes based on the type of crimes, 5) keeping spatial and temporal references, 6) reporting modus operandi and reason of the crimes, when present in the news articles, 7) finding correlations between different crime events.

2.2 Ontology design

The lightweight ontology is available online along with its documentation,[2] generated through WIDOCO (WIzard for DOCumenting Ontologies) [4] and is indexed in Linked Open Vocabularies. For a sound design practice, we also consider ontology design patterns related to the events, like EventCore[3] and NewsReportingEvent.[4] CEM has been designed to be event-centric, its structure is depicted in Figure 1[5] with 50 classes (18 integrated from existing sources), 10 datatype properties (3 from existing sources) and 21 object properties (16 from existing sources).

The main class is `sem:Core` that has 4 sub-classes: `sem:Event` for generic event representation, `sem:Place` for event localization, `sem:Time` for temporal reference, and `sem:Actor` for people participating in the event. The `sem:Event` class has been specialized with a new class named `cem:CrimeEvent` that has several sub-classes representing the different types of crime. The classification of crimes is derived from the International Classification of Crime for Statistical Purposes (ICCS),[6] an international standard endorsed by the United Nations Statistical Commission and the Commission on Crime Prevention and Criminal Justice in 2015 for the collection and statistical analysis of crime data. This classification does not care about the legal representation of crimes, but rather it provides a description of what happened to make data from different countries comparable. Since the classification is more detailed than the description usually provided by the newspapers, only the first two levels of the hierarchy have been taken into account, with just one exception for the `cem:MoneyLaundering` class. In total, there are 28 classes for the specification of the crime type: 11 classes in the first level of the hierarchy, 16 in the second level and 1 in the third level. however, other classes can be easily included.

[2] http://crimeanalytics.it/cem-reference/doc/index-en.html

[3] http://krisnadhi.github.io/onto/event.owl

[4] http://semantic.cs.put.poznan.pl/ontologies/newsreportingevent.owl

[5] The zoomable figure is available at http://crimeanalytics.it/cem-reference/doc/cem-model-full.png.

[6] https://www.unodc.org/unodc/en/data-and-analysis/statistics/iccs.html

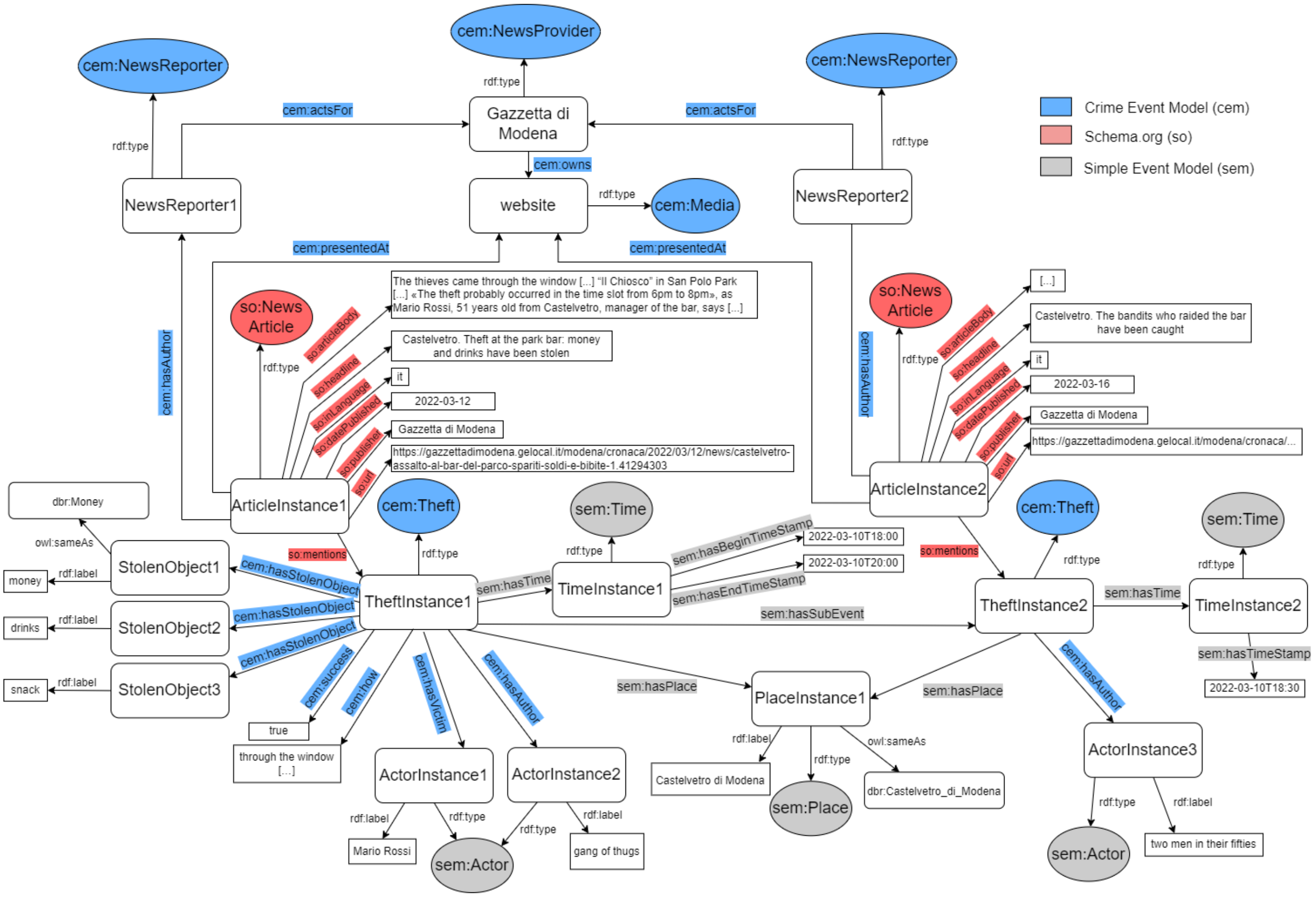

Figure 2: CEM usage example.

New properties have been defined: cem:hasAuthor, cem:hasVictim and cem:hasOtherSubject with domain cem:CrimeEvent and range sem:Actor to link the crime event to the author(s), victim(s) and other people involved in the event (e.g., LEAs, bystanders witnessing the crime), respectively.

Other new relevant properties are: cem:success to indicate if the crime has been successful or not; the allowed values are boolean, cem:why to express the crime motivation as reported in the news article, cem:how to identify how the crime has been accomplished. Since cem:why and cem:how can hold free text we choose to set the range of those properties as strings. Additional properties are provided for specific categories of crime, e.g., cem:hasStolenObject to indicate the objects that have been stolen during a theft.

Since the crime-related data are extracted from online newspapers, the so:mentions property of the *Schema.org* ontology is exploited to keep the reference to the online news articles. In particular, the class so:NewsArticle with the properties to specify article headline, language, body, url, and date of publication have been integrated. Information about the publisher are integrated considering the approach used in NewsReportingEvent. In particular, we exploit three classes that are connected to so:NewsArticle: NewsReporter indicates the author of the news article, NewsProvider could be the name of the newspaper, and Media is the mean for publication (e.g., website, Facebook, Twitter, and other social networks). This choice of design allows to integrate news articles of the same provider published on different channels and compare them.

In the end, the property sem:hasSubEvent from SEM is used to link together follow-up news articles. Indeed, in journalism, it is very typical to publish news to provide updates about an event.

The CEM lightweight ontology does not provide a classification of the property objects; however, the instances of each property (e.g., places, actors, stolen objects) can be linked to external elements of well-known ontologies, taxonomies or thesauri, e.g., WordNet, Wikidata, DBpedia, LinkedGeoData.

3 EXEMPLAR CRIME EVENT ANNOTATION

CEM can be used on crime news streams that are already classified, or that can be classified using supervised or unsupervised methods [3, 10]. An example of CEM usage is given in Figure 2 that illustrates two news articles instances referring to the same crime event (cem:Theft) happened in a city named Castelvetro di Modena. The news articles are taken from the Italian Crime News dataset[7] [11] which collects some news articles from the "Gazzetta di Modena" newspaper and extracts semantic information from the news articles' body. In particular, one of the two news articles in the Figure derives from the translation from Italian to English of news published in March 2022.[8]

The two news articles are related to a theft, and the second news, published on the 16th March, provides updates about the author of the theft w.r.t. the previous news published on the 12th March. Indeed, the two instances of `cem:Theft` are linked each other via the `sem:hasSubEvent` property and the second instance is linked to another instance of `sem:Actor` to explain the authors of the theft. As can be noticed, integration with DBpedia is also provided in two cases: `dbr:Money` and `dbr:Castelvetro_di_Modena`. The same approach could be used to integrate other instances also for other properties and with other vocabularies. For example, snack and drinks could be linked via WordNet to their shared hypernym "nutrient". In this way, it is possible to make analysis on the thefts of a specific class of objects. Semantic networks like WordNet, ConceptNet or BabelNet are suitable to build a classification of objects/entities, while Wikidata can be exploited to export a hierarchy of municipalities, province, areas and countries from all over the World to detect hotspots in crime data.

4 CONCLUSION AND FUTURE WORK

In this paper, a new lightweight ontology named Crime Event Model, CEM, has been introduced. The main purpose of development is supporting crime analysis with data extracted from newspapers. The model is event-centric and is designed to represent crime events as they are described in news articles. CEM is available online and is compliant with the FAIR principles of making the ontology findable, accessible, interoperable and reusable.[9] A classification of crime is provided, however, the model is easily extendable to include additional types of crime. Since the proposed lightweight ontology does not provide a legal representation (i.e., a description according to the criminal law) it can be used in any country. Such a model was not yet present in literature as the existing ontologies are not focused on crimes or they provide a legal representation of them. The last case is out of the scope of the paper.

In previous works, we used Knowledge Graphs without the support of ontologies to analyze news articles reporting crime events [12]. Further steps of our research will focus on the automatic population of the CEM ontology and the application of community detection algorithms [8] and pattern recognition approaches to demonstrate how the proposed lightweight ontology can be used to conduct event-based analysis.

ACKNOWLEDGMENTS

This work is partially supported by the project "Deep Learning for Urban Event Extraction from News and Social media streams" founded by the Engineering Department "Enzo Ferrari" of the University of Modena and Reggio Emilia.

REFERENCES

[1] Jans Aasman. 2020. *Entity Event Knowledge Graphs for Data Centric Organizations.* White Paper. Franz Inc. 7 pages. https://allegrograph.com/wp-content/uploads/2020/06/Entity-Event-Knowledge-Graphs-White-Paper-v692020.pdf

[2] Carmelo Asaro, Maria Angela Biasiotti, P. Guidotti, Maurizio Papini, Maria-Teresa Sagri, and Daniela Tiscornia. 2003. A Domain Ontology: Italian Crime Ontology. In *ICAIL Workshop on Legal Ontologies & Web Based Legal Information Management.*

[3] Giovanni Bonisoli, Federica Rollo, and Laura Po. 2021. Using Word Embeddings for Italian Crime News Categorization. In *Proceedings of the 16th Conference on Computer Science and Intelligence Systems, Online, September 2-5, 2021 (Annals of Computer Science and Information Systems, Vol. 25)*, Maria Ganzha, Leszek A. Maciaszek, Marcin Paprzycki, and Dominik Slezak (Eds.). 461–470. https://doi.org/10.15439/2021F118

[4] Daniel Garijo. 2017. WIDOCO: a wizard for documenting ontologies. In *International Semantic Web Conference.* Springer, Cham, 94–102. https://doi.org/10.1007/978-3-319-68204-4_9

[5] Daniel Garijo, Óscar Corcho, and María Poveda-Villalón. 2021. FOOPS!: An Ontology Pitfall Scanner for the FAIR principles. In *Proceedings of the ISWC 2021 Posters, Demos and Industry Tracks held with 20th International Semantic Web Conference (ISWC 2021), online, October 24-28, 2021 (CEUR Workshop Proceedings, Vol. 2980)*, Oshani Seneviratne, Catia Pesquita, Juan Sequeda, and Lorena Etcheverry (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-2980/paper321.pdf

[6] Simon Gottschalk and Elena Demidova. 2019. EventKG - the hub of event knowledge on the web - and biographical timeline generation. *Semantic Web* 10, 6 (2019), 1039–1070. https://doi.org/10.3233/SW-190355

[7] Simon Gottschalk, Endri Kacupaj, Sara Abdollahi, Diego Alves, Gabriel Amaral, Elisavet Koutsiana, Tin Kuculo, Daniela Major, Caio Mello, Gullal S. Cheema, Abdul Sittar, Swati, Golsa Tahmasebzadeh, and Gaurish Thakkar. 2021. OEKG: The Open Event Knowledge Graph. In *Proceedings of the 2nd International Workshop on Cross-lingual Event-centric Open Analytics co-located with the 30th The Web Conference (WWW 2021), online, April 12, 2021 (CEUR Workshop Proceedings, Vol. 2829)*, Elena Demidova, Sherzod Hakimov, Jane Winters, and Marko Tadic (Eds.). CEUR-WS.org, 61–75. http://ceur-ws.org/Vol-2829/paper5.pdf

[8] Laura Po and Davide Malvezzi. 2018. Community Detection Applied on Big Linked Data. *J. Univers. Comput. Sci.* 24, 11 (2018), 1627–1650. http://www.jucs.org/jucs_24_11/community_detection_applied_on

[9] María Poveda-Villalón, Paola Espinoza-Arias, Daniel Garijo, and Óscar Corcho. 2020. Coming to Terms with FAIR Ontologies. In *Knowledge Engineering and Knowledge Management - 22nd International Conference, EKAW 2020, Bolzano, Italy, September 16-20, 2020, Proceedings (Lecture Notes in Computer Science, Vol. 12387)*, C. Maria Keet and Michel Dumontier (Eds.). Springer, 255–270. https://doi.org/10.1007/978-3-030-61244-3_18

[10] Federica Rollo, Giovanni Bonisoli, and Laura Po. 2021. Supervised and Unsupervised Categorization of an Imbalanced Italian Crime News Dataset. In *Information Technology for Management: Business and Social Issues - 16th Conference, ISM 2021, and FedCSIS-AIST 2021 Track, Held as Part of FedCSIS 2021, Virtual Event, September 2-5, 2021, Extended and Revised Selected Papers (Lecture Notes in Business Information Processing, Vol. 442)*, Ewa Ziemba and Witold Chmielarz (Eds.). Springer, 117–139. https://doi.org/10.1007/978-3-030-98997-2_6

[11] Federica Rollo and Laura Po. 2020. Crime Event Localization and Deduplication. In *The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12507)*, Jeff Z. Pan, Valentina A. M. Tamma, Claudia d'Amato, Krzysztof Janowicz, Bo Fu, Axel Polleres, Oshani Seneviratne, and Lalana Kagal (Eds.). Springer, 361–377. https://doi.org/10.1007/978-3-030-62466-8_23

[12] Federica Rollo and Laura Po. 2022. Knowledge Graphs for Community Detection in Textual Data. In *Knowledge Graphs and Semantic Web - 4th Iberoamerican Conference and third Indo-American Conference, KGSWC 2022, Madrid, Spain, November 21-23, 2022, Proceedings (Communications in Computer and Information Science, Vol. 1686)*, Boris Villazón-Terrazas, Fernando Ortiz-Rodríguez, Sanju Tiwari, Miguel-Ángel Sicilia, and David Martín-Moncunill (Eds.). Springer, 201–215. https://doi.org/10.1007/978-3-031-21422-6_15

[13] Willem Robert van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. 2011. Design and use of the Simple Event Model (SEM). *J. Web Semant.* 9, 2 (2011), 128–136. https://doi.org/10.1016/j.websem.2011.03.003

[14] Mark D. et al. Wilkinson. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 1 (2016), 160018. https://doi.org/10.1038/sdata.2016.18

[7] https://paperswithcode.com/dataset/italian-crime-news

[8] https://www.gazzettadimodena.it/modena/cronaca/2022/03/12/news/castelvetro-assalto-al-bar-del-parco-spariti-soldi-e-bibite-1.41294303

[9] The FAIR compliance was checked by FOOPS! (Ontlogy Pitfall Scanner for FAIR) [5].