A Comprehensive Extraction of Relevant Real-World-Event Qualifiers for Semantic Search Engines

Guillaume Bernard, Cyrille Suire, Cyril Faucher, Antoine Doucet

A Comprehensive Extraction of Relevant Real-World-Event Qualifiers for Semantic Search Engines

Guillaume Bernard[1 [0000−0001−5945−4865]], Cyrille Suire[1], Cyril Faucher[1], and Antoine Doucet[1 [0000−0001−6160−3356]]

Université de La Rochelle, Laboratoire L3i, 17000 La Rochelle, France
{guillaume.bernard,cyrille.suire,cyril.faucher,antoine.doucet}@univ-lr.fr
https://l3i.univ-larochelle.fr/

Abstract. In this paper, we present an efficient and accurate method to represent events from numerous public sources, such as Wikidata or more specific knowledge bases. We focus on events happening in the real world, such as festivals or assassinations. Our method merges knowledge from Wikidata and Wikipedia article summaries to gather entities involved in events, dates, types and labels. This event characterization procedure is extended by including vernacular languages. Our method is evaluated by a comparative experiment on two datasets that shows that events are represented more accurately and exhaustively with vernacular languages. This can help to extend the research that mainly exploits hub languages, or biggest language editions of Wikipedia. This method and the tool we release will for instance enhance event-centered semantic search engines, a context in which we already use it. An additional contribution of this paper is the public release of the source code of the tool, as well as the corresponding datasets.

Keywords: Event · Information Retrieval · Linked and Open Data.

1 Introduction

Analysing and characterising events in natural language processing has multiple applications. One of them is semantic search engines, used to browse large digital libraries or press articles [1,2,23]. To build an efficient event based query, the event representation and description are crucial. It is necessary to collect exhaustive information from data sources in order to be as precise as possible when qualifying events. This means being able to answer some simple questions, such as where the event happened, and when and who or what was involved [5,24]. The answers to these questions are often named entities [32], and considered as event qualifiers. To the best of our knowledge, the state of the art is missing a method to extract an as comprehensive as possible representation of real-world-events. Research projects often propose their own definition of events and adopt their own representation that fulfill their needs. Reference data sources are numerous and nothing exists to exploit them in a unified way.

The purpose of this paper is to overcome these limitations by taking advantage of past experiments and to provide an efficient representation of events by addressing two different issues. We first wish to know how to qualify, in any language, real-world-events based on publicly available resources. Subsequently, we propose an approach to obtain an almost comprehensive event qualification by exploiting vernacular languages, that is to say, the languages spoken where the events happened. We try to demonstrate the most spoken languages are not sufficient to accurately extract event qualifiers and that vernacular languages must be processed as well.

2 Related Work

In the recent years, a lot of publicly available data sources emerged to provide a universal access to multilingual information. A lot of them took benefit from Wikipedia, the largest knowledge base in the history of human kind. Through the years, many ontologies projects aimed at extracting the semantic knowledge of Wikipedia articles. Back in 2007, the DBPedia project [15] was the first knowledge graph (KG) to gather data from Wikipedia articles, infoboxes and lists. Released a few years after, YAGO2 [11] inherits the same characteristics. It is built from Wikipedia and supplemented with WordNet [18] and GeoNames information. YAGO2 is linked to DBPedia entities. A year after, the Wikimedia Foundation unveiled Wikidata [30], a community maintained knowledge graph. This one is used as a reference graph to harmonize content across versions of Wikipedia. Recent projects investigate automatic writing of Wikipedia articles in low endowed Wikipedia linguistic versions [29]. The AbstractWikipedia project aims at solving an automated text generation task from semantic knowledge hosted on Wikidata. This is made possible as Wikidata is one of the highest qualitative multilingual knowledge repository, even if the amount and quality of its knowledge is not always connected to the number of worldwide native speakers [12]. A survey [6] compared these knowledge graphs in terms of quality according to many metrics and gives criteria to find the most suitable graph for the needs of researchers.

None of these graphs is dedicated to events. EventKG [10] fills this gap. It is based on the Simple Event Model [28] (SEM) ontology and intended to store events. SEM focuses on events elementary characteristics: types, dates, locations and participants. It defines a simple model to represent real-world-events. The role of entities associated to events is to point out who, what, where and when the event happened [31]. EventKG aggregates data from multiple sources and connects them in a graph to ensure easy communication through the semantic web. EventKG is good to represent major events such as happenings, festivals and disasters [20].

Characterizing real world events has become an important research issue for a few years. While a lot of work has been made to detect and extract events from news articles [14,16], another trend consists of extracting the semantic knowledge from data sources in order to connect real world events to news and press articles.

This purpose is striven towards the usage for digital libraries by providing for instance semantic or event based search engines to explore historical news [23]. Wikipedia articles lead sections offer qualitative data and give an overall picture of events [20]. They often contain elementary event information: dates, places, and participating entities. The latter have also been used to extract events from real time news, with a particular focus on people, organisations, places and dates when semantically enriching documents [13]. Semantic labeling thanks to Wikidata and Wikipedia [5] has been proven useful in semantic search engines [21] with annotated press articles.

We notice in the state of the art that exploiting Wikidata and Wikipedia entities and elementary event knowledge is useful, especially with specific use cases as exploring digital libraries. We propose to associate Wikidata and Wikipedia to exhaustively collect real-world-event qualifiers which are dates, places and participating entities. Ontologies such as EventKG do not provide comprehensive data, some entities may miss. We propose to collect knowledge where it is: from encyclopedias. On another hand, we know Wikimedia projects are multilingual. Some studies on event mentions tracking suggested to focus on hub languages [22] (languages with a high number of articles and significant overlap in article coverage) to qualify events, we will propose another approach, based on vernacular languages.

3 Representing Events from Wikidata and Wikipedia

At first, we address the first question: we wish to qualify, in any language, real-world-events based on publicly available resources. We present the method we developed to collect event qualifiers from Wikidata and Wikipedia. We consider they both provide sufficient data to characterize real world events. We act in the continuation of the Automated Content Extraction program [3] and existing event ontologies [28,24]. Our event qualifiers are used in the same context as ACE's event arguments. We benefit from Wikidata to extract elementary event information and use Wikipedia to aggregate all the entities involved in it.

3.1 The Extraction of Elementary Event Information

Wikidata supplies two different event identifiers, which point subtleties: some apply for breaking events, others for event with premisses. In this paper, we conform to the Wikidata Event Type (WET) [21] definition which accepts both. We refer to events as *happenings in the real world which have spatio-temporal anchors and additional entities involved in it.* From Wikidata entities, we only collect the event type and date, the locations, participants and labels. We consider these properties discriminate two events: it is unlikely that two distinct events have the same label, type, and occurred at the same place at the same time. As a community project, Wikidata is not an exhaustive data source. Table 1 shows that expecting a comprehensive event qualifiers collection is not possible when only capitalizing on Wikidata. For instance, almost all events miss the *involved participant* property.

Table 1. The proportion of WETs with location, date and participants qualifiers. There is a total of 952.351 events.

Named entities category [27]	Wikidata property	Number of events	Percentage
PER[SON]	Participant (*P710*)	58,885	6.18%
DATE	Time (*P585*, *P580*, *P582*)	511,312	53.69%
LOC[ATION]	Location (*P7*, *P276*)	524.532	55.08%

Table 2. Properties of the different language editions of Wikipedia. Sorted by decreasing number of articles (*Data collected in Oct. 2020*).

Language	Articles	Modified pages	Contributors	Active Contributors	Article Depth
	in millions			*in thousands*	
English	6.151	1200	386	32	1026.81
German	2.475	241	50	5.5	93.6
French	2.246	280	54	5.1	237.67
Russian	1.657	173	40	3.4	135.94
Italian	1.631	183	44	2.5	169.03
Spanish	1.622	270	87	4.2	208.81
Polish	1425	113	14	1.3	30.99

3.2 Entities Involved in the Event

To go beyond, we propose to analyze Wikipedia lead sections in search of participating entities. A lead section (i.e., a summary) on Wikipedia contains a lot of important information, a synthesis of the article itself and reports the main topic [9]. On Wikipedia, internal links connect articles to Wikidata. We use Wikipedia lead section internal links to detect entities involved in the event. We assume it is possible to add time, location and participant information, when they are absent from Wikidata.

There are, in April 2021, 310 active language editions of Wikipedia. First, and in the interest of efficiency, we presume we can only focus on some languages with the most articles, hub languages (Table 2). In addition to the number of articles, we included the number of modified pages, of contributors (and active ones) and the Wikipedia article depth [8]. The latter is a Wikipedia article quality indicator based on content edits. From the top-ten language list, Cebuano (2^{nd}), Swedish (3^{rd}) and Dutch (6^{th}) are mainly bot written and therefore excluded. In case of conflict, the higher Wikipedia depth, the higher priority. We decided to arbitrarily select five languages. Following the criteria mentioned earlier, we kept English, German, French, Italian and Spanish, covering native languages of 30% of the world's population [4] and 25.14% of all Wikipedia articles (14.125 [7] over 56.615 million articles). We suppose this set of Wikipedia editions is sufficient to accurately gather event qualifiers. Nevertheless, this selection is biased and excludes most Asian and African languages with large speaker communities, as Mandarin and Hindi, whose Wikipedia versions are smaller. They respectively gather 1.120 billion and 128 thousands of articles [7].

From articles lead sections, we keep people, locations and organisations or geopolitical entities. The number of occurrences found for each entity in all lead sections is counted and represents the entity weight in relation to the event. This weight shows the relevance of the entity in relation to the event. We assume the entities found in multiple lead sections are important entities in the event description.

Let us take the example of the *assassination of Rasputin* event on Wikidata (identified by *Q2882749*). From Wikidata, we retain the date, locations, participants and entity labels in multiple languages. We ignore, for the time being, other properties associated with the event type, such as the *target* for a *political assassination*. Participants and locations are linked entities and identified by their URIs in the ontology. We supplement the event characterization with entities found in Wikipedia articles. There only exists Wikipedia articles written in French and Spanish for this event. After processing, we obtain, among others, these triples: *(PER, Q312997 [Felix Yusupov, perpetrator], 3)*, *(PER, Q43989 [Grigori Rasputin, target], 2)*, *(PER, Q34266 [Stanislas Lazovert], 1)*. The weights, respectively 3, 2 and 1 show that *Yusupov* is a major player in the event, while *Lavozert* has a limited implication, even if he is a known plotter. Weights synthesize historical knowledge and give an unbiased information about entities implication in the event. The *Lavozert* entity is absent from Wikidata and was extracted from the French lead section.

3.3 Localizing Event Qualifiers

The event description consists of an association of absolute properties such as the date and labels with links to knowledge bases. The description is fundamentally multi-lingual. In most cases, Wikidata provides multiple names in different languages (*i.e.* with different spellings) for each entity. To continue with the previous example, in French the entity *Q312997* on Wikidata is equally written *Félix Youssoupoff* or *Felix Youssoupov*.

This final step transforms abstract entities, identified by their Wikidata URIs to a language-dependent description. It takes all the alternative spellings for every entity involved in the event and saves them in the targeted language. Our approach makes it possible to get the event description in Italian even if, in this example, only French and Spanish Wikipedias were analyzed.

3.4 Conclusion

In this section, we proposed a method to characterize real-world events with qualifiers. Our method relies on the Wikidata ontology and Wikipedia to extract all the event participating entities. Our approach is multi-lingual: entities are identified by URIs but can be turned into any existing language. By selecting a subset of all the available Wikipedia languages, we assume we can efficiently collect most of the event entities. With this paper, we release the `wikivents` tool[1] that implements the method described in this section. It is able to automatically extract the event representation and participating entities with a Wikidata identifier as input. The package can be customized in order to gather data from other resources, out of Wikimedia projects. More information about its API and tutorials are available in the project archive.

4 Enhancing the Event Representation with Vernacular Languages

Although we introduced in Section 3 our method to collect most of the event qualifiers, the arbitrary selection of some hub languages is biased. Widely spoken languages that are less present on Wikipedia are ignored. Numerous languages (*i.e.* Arabic, Mandarin, Hindi, Bengali, Portuguese or Russian) are concerned. We intend to discover whether the language influences the event representation when processing Wikipedia articles. To answer this second question, we propose to extend the list of processed languages with vernacular languages.

In this section, we carry out a comparative experiment to know how useful and pertinent it is to benefit from vernacular language when processing Wikipedia articles. First, we introduce the dataset we built, then the evaluation process and our results. We conclude with a short error analysis.

4.1 Datasets description

Selected events In order to compare the influence of language, we built two distinct datasets with the same events. As the nature of an event is ambiguous [26], we qualify of indisputable an event that is considered as such for people with various backgrounds (history scholars [25], psychologists [19] or NLP reseachers [17], for instance). This led us to restrict events to only three categories, taken as examples: assassinations and attacks, natural disasters and political happenings. On Wikidata, the first two concern breaking events while the last gathers events with premisses. We randomly selected two event types for each category.

- **Assassination and attacks**: political murder (*Q1139665*) and terrorist attack (*Q2223653*).
- **Natural disasters**: earthquake (*Q7944*) and volcanic eruption (*Q7692360*).
- **Political happenings**: ceremony (*Q2627975*) and election (*Q40231*).

We express the same reserve about the exhaustiveness of Wikidata. The number of events reported is not uniform over the years but tends to grow since the beginning of the 21st century. This increase must not be interpreted as an increase of events happening in the world but as a better data quality, especially with events now better anchored in time [21]. Therefore, we decided

[1] The package is a Python 3 library called `wikivents` on Pypi.org and available on the Software Heritage repository at `https://archive.softwareheritage.org/swh:1:dir:ef325a054ba6f7eb1121807da7b1c92b9ecde8f8`

Table 3. The number and ratio of events with at least one Wikipedia article in any language (even not a hub language), from 1970 to 2019.

Category	Event type	Events	With an article	Ratio
Assassinations and attacks	Political murder	44	24	54.55%
	Terrorist attack	905	806	89.06%
Natural disasters	Earthquake	1,102	987	89.56%
	Volcanic eruption	23	18	78.26%
Political happenings	Ceremony	11,428	11,233	98.29%
	Election	29,236	24.488	82.10%

to only focus on events happening in the last fifty years, from January 1970 to December 2019. This ensures to exclude poorly documented events. For the sake of the experiment, it is necessary to process events which are described in at least one Wikipedia article. Thus, we exclude events without any articles. The number of found events is reported in Table 3.

The process of gathering participating entities described in Section 3 may be slow for some events. It produces numerous API calls and can take from a few seconds to a significant amount of time depending on the number of lead sections to be processed. Consequently, we randomly select a maximum of fifty events for every event type. Selected events all satisfy the previously mentioned requirements: they have a date property and at least one related article.

Vernacular languages The difference between the two datasets resides in the number of languages processed to gather participating entities. The first one is called “base language” and built from the five languages mentioned above. We call the other one “all languages” for which vernacular languages are processed in addition to those from the base dataset. In case a Wikipedia project in this language is missing, we fall back to its standard dialect. For instance, American English is “en-us” which does not exist on Wikipedia, but is a dialect of “en”.

This process introduces another bias: temporality. French was an official language in Algeria before 1950. This information is missing from Wikidata which only provides current information. Moreover, the ontology sometimes provides the languages spoken in the countries, including the non official ones. We state the hypothesis that if an event occurs somewhere, the event will be better reported on Wikipedia in one of the languages spoken where it happened.

It may happen that the political entity changed through the years: the Easter Rising occurred in the *United Kingdom of Great Britain and Ireland.* This country does no longer exist, as the current *United Kingdom* exists since 1922. This is not an issue as Wikidata still informs about official or spoken languages. In any case, our dataset comprises events from 1970 to 2019, which limits this risk.

Table 4. The four metrics for the selected languages about the assassination of JFK.

	Language				
Characteristic	Italian	Spanish	German	English	French
M1: participating entities in the lead section	40	42	13	**47**	35
M2: tokens found in the lead section	218	125	179	**316**	279
M3: ratio between the two previous metrics	0.183	**0.336**	0.073	0.149	0.125
M4: alternative names for each Wikidata entity	83	130	166	**224**	115

Conclusion With this paper, we release the datasets we built. [2] [3] They comprise 241 events divided up among six events types, in three categories. The first dataset is produced with selected languages, the other with selected and vernacular languages. Events were processed using the `wikivents` tool we described in Section 3. For each of them, we also provide the content of the lead section that was used to extract entities and we saved each event in all the processed languages to simplify any further analysis.

4.2 Experimental Metrics

The comparison between the two datasets is based on four metrics we consider as relevant to evaluate our hypothesis: processing vernacular languages on Wikipedia supplies additional and more precise information about participating entities. Metrics respectively are the number of participating entities found in the lead section of Wikipedia articles (**M1**), the number of tokens in the lead section, to the exclusion of tokens shorter than two characters (**M2**), the ratio between the two previous metrics (**M3**) and the number of alternative names found for each Wikidata entity, as mentioned in Section 3.3 (**M4**).

Table 4 records the results for the assassination of John Fitzgerald Kennedy in 1963. As the assassination took place in the USA, an English speaking country, no additional language gets processed in addition to the five default ones. We notice the English version of Wikipedia provides more information about this event. More entities are found in the lead section, the lead section length is longer and we have more alternative names in English than in order languages.

In a second phase, languages are sorted in decreasing order, the higher value in Table 4, the first. Sorts for this events are shown in Table 5 and demonstrate the English version of Wikipedia is the most accurate to describe the event participating entities. English is in first place in terms of participating entities, tokens and number of alternative names found. The ratio is sometimes erroneous due to the different writing styles adopted by the multiple language Wikipedia communities. This may explain why Spanish is in first place for this metric.

[2] The dataset is hosted on Zenodo: `https://doi.org/10.5281/zenodo.4733506`.

[3] Available on the Software Heritage repository at `https://archive.softwareheritage.org/swh:1:dir:ef325a054ba6f7eb1121807da7b1c92b9ecde8f8`

Table 5. Ranked languages that best represent the assassination of JFK

Characteristic	1st	2nd	3rd	4th	5th
Participating entities	**en**	es	it	fr	de
Tokens in the lead section	**en**	fr	it	de	es
Ratio	es	it	**en**	fr	de
Alternative names	**en**	de	es	fr	it

Table 6. Number of events excluded because of a missing spoken language, or because of missing Wikipedia articles.

Event type	Events	Without spoken language given	Without article	
			Base language	*All languages*
Political murder	24	0	1	0
Terrorist attack	50	13	5	2
Earthquake	50	11	17	11
Volcanic eruption	17	1	3	2
Ceremony	50	29	9	6
Election	50	10	11	2

4.3 Experiment Evaluation

Although when building the dataset in Section 4.1 we excluded events without any Wikipedia article, it may happen that some selected events do not exist on Wikipedia in any of the processed languages. They were not filtered out in the first step because they have at least an article, but written in a language which is not a hub or a vernacular language, which we were unaware of at the first step. It is also a necessity to exclude events for which we do not know either any official or spoken language. The two overlap in most cases. We report in Table 6 the number of events excluded by this final selection. When considering vernacular languages, the number of events to analyze increases. This is the first argument in favour of our hypothesis that events are better described in their vernacular languages.

In order to compare the description of events in the two datasets, we apply, for each event, the same computations as shown in Table 4 and Table 5. For each metric, we check whether one of the official languages spoken in the event place is in the best three languages to characterize it. Results shown in Table 7 show, for each metric, for how many events a vernacular language is in the top three of language that best represent the event. Results are significant with only the first language but selecting the best third languages tends to limit the issue described with the Kennedy's assassination example. By doing so, we state that the five languages we previously identified as core languages are not sufficient to accurately extract event qualifiers. This statement refutes the hypothesis we assumed in Section 3 that led us to only consider only five hub languages.

Table 7. Comparison of how many events are better described by a vernacular language. The top-three languages that best represent the event are taken into account.

Event type	Dataset	Number of events	Events better described by a vernacular language			
			M1: entities	*M2: tokens*	*M3: ratio*	*M4: alt. names*
Political murder	Base	23	14	14	13	12
	All		**22**	**22**	**20**	**18**
Terrorist attacks	Base	34	25	25	25	27
	All		**30**	**30**	**29**	**28**
Earthquake	Base	25	13	13	12	11
	All		**24**	**24**	**23**	**20**
Volcanic eruption	Base	14	12	12	12	12
	All		**13**	**13**	**13**	**13**
Ceremony	Base	18	12	13	13	15
	All		**15**	**15**	**15**	**17**
Election	Base	31	27	27	26	26
	All		**30**	**30**	**29**	26

4.4 Error Analysis

For the majority of events that contradict the hypothesis, the main reason is a lack of resources in the vernacular languages: we miss Wikipedia articles so cannot extract any lead section. Missing articles are due to a small community of speakers and then results in a small Wikipedia edition or may be explained by cultural bias. The latter is mainly true for assassination and attacks events which are treated, or not, differently in the Wikipedia language editions. For few events, the vernacular language is in fourth position or even further away and is not the best at representing the given event.

We mainly processed Indo-European languages for which the tokenization procedure is quite uniform, then comparable. This is a noticeable limit of our analysis regarding some of our metrics.

5 Conclusion

In this paper, we described a method to gather event qualifiers coming from Wikidata and Wikipedia. We analysed and described its shortcomings and proposed to include vernacular languages. Our experiments demonstrate that this approach is greatly beneficial when describing events. We release an implementation of our approaches, and actually hereby make publicly available the source code, the analysis as well as the datasets, to be updated regularly. In the near future, we will add features to encode the event model into SEM [28] or LODE [24]. Researchers working on real-world events may already take advantage of our tool to fulfil their needs. It can be used to qualify events for their semantic search engines. We already use the library as the entry point of an event based search engine for a historical news digital library. It uses the event representation from the `wikivents` library in order to forge queries to retrieve documents associated to events [**?**].

6 Acknowledgments

This work has been supported by the European Union's Horizon 2020 research and innovation program under grants 770299 (NewsEye).

References

1. Brank, J., Leban, G., Grobelnik, M.: Semantic Annotation of Documents. Informatica **42**, 23–32 (Jan 2017)
2. Cybulska, A.K., Vossen, P.: Historical Event Extraction from Text. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp. 39–43. Portland, Oregon, USA (Jun 2011), https://www.aclweb.org/anthology/W11-1506
3. Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R.: The Automatic Content Extraction (ACE) program. Tasks, Data and Evaluation. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). pp. 837–840. Lisbon, Portugal (May 2004), http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf
4. Eberhard, David M., Gary F. Simons, Charles D. Fennig: Ethnologue: Languages of the World (2021), https://www.ethnologue.com/
5. Exner, P., Nugues, P.: Using semantic role labeling to extract events from wikipedia. In: DeRiVE@ ISWC. pp. 38–47 (2011)
6. Färber, M., Bartscherer, F., Menne, C., Rettinger, A.: Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. Semantic Web **9**(1), 77–129 (Nov 2017). https://doi.org/10.3233/SW-170275
7. Foundation, T.W.: List of Wikipedias. Wikipedia (Apr 2021), https://en.wikipedia.org/w/index.php?title=List_of_Wikipedias&oldid=1016309550
8. Foundation, T.W.: Wikipedia article depth - Meta (Apr 2021), https://meta.wikimedia.org/wiki/Wikipedia_article_depth
9. Foundation, T.W.: Wikipedia:Summary style. Wikipedia (Apr 2021), https://en.wikipedia.org/w/index.php?title=Wikipedia:Summary_style&oldid=1015628666
10. Gottschalk, S., Demidova, E.: EventKG - the Hub of Event Knowledge on the Web - and Biographical Timeline Generation. Semantic Web **10**(6), 1039–1070 (Oct 2019). https://doi.org/10.3233/SW-190355
11. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence **194**, 28–61 (Jan 2013). https://doi.org/10.1016/j.artint.2012.06.001
12. Kaffee, L.A., Piscopo, A., Vougiouklis, P., Simperl, E., Carr, L., Pintscher, L.: A Glimpse into Babel: An Analysis of Multilinguality in Wikidata. In: Proceedings of the 13th International Symposium on Open Collaboration - OpenSym '17. pp. 1–5. Galway, Ireland (2017). https://doi.org/10.1145/3125433.3125465
13. La Fleur, A., Teymourian, K., Paschke, A.: Complex event extraction from real-time news streams. In: Proceedings of the 11th International Conference on Semantic Systems. pp. 9–16. Vienna Austria (Sep 2015). https://doi.org/10.1145/2814864.2814870
14. Leban, G., Fortuna, B., Brank, J., Grobelnik, M.: Event registry: Learning about world events from news. In: Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion. pp. 107–110. Seoul, Korea (2014). https://doi.org/10.1145/2567948.2577024

15. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web **6**(2), 167–195 (2015). https://doi.org/10.3233/SW-140134
16. Mele, I., Bahrainian, S.A., Crestani, F.: Event mining and timeliness analysis from heterogeneous news streams. Information Processing & Management **56**(3), 969–993 (May 2019). https://doi.org/10.1016/j.ipm.2019.02.003
17. Mele, I., Crestani, F.: A Multi-Source Collection of Event-Labeled News Documents. In: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval - ICTIR '19. pp. 205–208. Santa Clara, CA, USA (2019). https://doi.org/10.1145/3341981.3344253
18. Miller, G.A.: WordNet: A lexical database for English. Communications of the ACM **38**(11), 3 (Nov 1995)
19. Minsky, M.: A framework for representing knowledge. The Psychology of Computer Vision (1975)
20. Mishra, A., Berberich, K.: EXPOSÉ: EXploring Past news fOr Seminal Events. In: Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion. pp. 223–226. Florence, Italy (2015). https://doi.org/10.1145/2740908.2742844
21. Rudnik, C., Ehrhart, T., Ferret, O., Teyssou, D., Troncy, R., Tannier, X.: Searching News Articles Using an Event Knowledge Graph Leveraged by Wikidata. In: Companion Proceedings of The 2019 World Wide Web Conference on - WWW '19. pp. 1232–1239. San Francisco, USA (2019). https://doi.org/10.1145/3308560.3316761
22. Rupnik, J., Muhic, A., Leban, G., Skraba, P., Fortuna, B., Grobelnik, M.: News Across Languages - Cross-Lingual Document Similarity and Event Tracking. Journal of Artificial Intelligence Research **55**, 283–316 (Jan 2016). https://doi.org/10.1613/jair.4780
23. Shaw, R.: A Semantic Tool for Historical Events. In: Proceedings of the The 1st Workshop on EVENTS: Definition, Detection, Coreference, and Representation. pp. 38–46. Atlanta, Georgia, USA (Jun 2013)
24. Shaw, R., Troncy, R., Hardman, L.: LODE: Linking Open Descriptions of Events. In: The Semantic Web, vol. 5926, pp. 153–167. Springer Berlin Heidelberg, Berlin, Heidelberg (2009). https://doi.org/10.1007/978-3-642-10871-6
25. Shaw, R.B.: Events and Periods as Concepts for Organizing Historical Knowledge. Ph.D. thesis, UC Berkeley (2010), `https://escholarship.org/uc/item/4111f1fw`
26. Sprugnoli, R.: Event Detection and Classification for the Digital Humanities. Ph.D. thesis, Università degli Studi di Trento, Trento, Italia (Apr 2018), `http://eprints-phd.biblio.unitn.it/2865/`
27. Sundheim, B.M.: Overview of results of the MUC-6 Evaluation. In: Proceedings of the 6th Conference on Message Understanding. pp. 13–31 (Nov 1995). https://doi.org/10.3115/1072399.1072402
28. van Hage, W.R., Malaisé, V., Segers, R., Hollink, L., Schreiber, G.: Design and use of the Simple Event Model (SEM). Journal of Web Semantics **9**(2), 128–136 (Jul 2011). https://doi.org/10.1016/j.websem.2011.03.003
29. Vrandečić, D.: Architecture for a multilingual Wikipedia. arXiv:2004.04733 [cs] (Apr 2020), `http://arxiv.org/abs/2004.04733`
30. Vrandečić, D., Krötzsch, M.: Wikidata: A free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (Sep 2014). https://doi.org/10.1145/2629489
31. Xiang, W., Wang, B.: A Survey of Event Extraction From Text. IEEE Access **7**, 173111–173137 (Nov 2019). https://doi.org/10.1109/ACCESS.2019.2956831

32. Yadav, V., Bethard, S.: A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. Proceedings of the 27th International Conference on Computational Linguistics p. 14 (Aug 2018)