This is a post-peer-review, pre-copyedit version of an article published in the Springer CCIS series.
The final authenticated version is available online at: https://doi.org/10.1007/978-3-030-11226-4_13

Populating Narratives Using Wikidata Events: An Initial Experiment

Daniele Metilli, Valentina Bartalesi, Carlo Meghini, and Nicola Aloia

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" – CNR, Pisa, Italy
{daniele.metilli,valentina.bartalesi,carlo.meghini,nicola.aloia}@isti.cnr.it

Abstract. The study presented in this paper is part of our research aimed at improving the search functionalities of current Digital Libraries using formal narratives. Narratives are intended as sequences of events. We present the results of an initial experiment to detect and extract implicit events from the Wikidata knowledge base in order to construct a narrative in a semi-automatic way. Wikidata contains many historical entities, but comparably few events. The reason is that most events in Wikidata are represented in an implicit way, e.g. by listing a date of birth instead of having an event of type "birth". For this reason, we decided to generate what we call the Wikidata Event Graph (WEG), i.e. the graph of implicit events found in Wikidata. We performed an initial experiment taking as case study the narrative of the life of Italian poet Dante Alighieri. Only one event of the life of Dante is explicitly represented in Wikidata as instance of the class *Q1190554 Occurrence*. Using the WEG, we were able to automatically detect 31 more events of Dante's life that were present in Wikidata in an implicit way.

Keywords: Wikidata · Narratives · Semantic Web · Ontology · Digital Libraries

1 Introduction

Currently, Digital Libraries (DLs) offer search functionalities that respond to a user's web-like query with a list of digital objects based only on their metadata descriptors. We believe that DLs should be able to provide *narratives* to their users in addition to lists of objects. We intend narratives as sequences of events defined by a narrator, endowed with factual aspects (who, what, where, when) and semantic relations. Narratives would allow DLs to provide more sophisticated information services to their users, going beyond the current state.

In order to introduce narratives in DLs, we developed an ontology to formally represent narratives [2], based on the CIDOC CRM standard [4]. Subsequently, we built a semi-automated Narrative Building and Visualising Tool (NBVT)[1], which allows the user to construct a narrative as a sequence of events. The tool has been used to construct four narratives[2] about different subjects: the life of

[1] https://dlnarratives.eu/tool.html
[2] https://dlnarratives.eu/narratives.html

Florentine poet Dante Alighieri, the life of Austrian painter Gustav Klimt, the history of the giant squid, and the history of climate change.

While building the narrative in NBVT, the user can add to each event some entities (e.g. people, places, objects) related to the subject of the narrative. These entities can be automatically imported from the Wikidata[3] knowledge base. Wikidata is a collaborative project hosted by the Wikimedia Foundation [11]. Containing more than 50 million entities, it is one of the largest general-purpose knowledge bases. In order to facilitate the user's work when building a narrative, we developed a mapping between our ontology and the Wikidata ontology which can be applied to import entities and events into our tool [1].

Unfortunately, the number of events (instances of the class *Q1190554 Occurrence*) contained in the knowledge base is relatively low, because Wikidata's ontology is not event-based. The knowledge about events is present in Wikidata, but it is generally represented in an implicit way. For instance, the birth of the Florentine poet Dante Alighieri is not represented as an event "Birth of Dante Alighieri", but instead the knowledge base contains a statement of the form "Dante Alighieri *place of birth* Florence" which directly links the poet to the city he was born in.

To solve this issue, we have decided to extract the events implicitly contained in the knowledge base, generating a graph that we call the Wikidata Event Graph (WEG). This graph can then be used to import events into NBVT in order to populate our ontology for narratives.

In the following, we present the generation of a subset of the WEG focused on events about people's lives, and an initial experiment to verify how much the events contained in it can improve the narrative building process. Section 2 describes our reasons for choosing Wikidata as reference knowledge base. Section 3 describes the extraction of the event graph from Wikidata. Section 4 describes the initial experiment about the narrative of Dante's life, and its results. Finally, Section 5 reports our conclusions and future works.

2 Wikidata as Reference Knowledge Base

In our formal representation, a narrative consists of three main elements: (i) *fabula*, i.e. the sequence of events in chronological order, (ii) *narration(s)*, i.e. one or more texts that express the narrative, and (iii) *reference function* that connects the narrations to the fabula, allowing the derivation of the *plot*. In the fabula, each event is endowed with entities such as people, places, and physical objects. When representing such events and entities through Semantic Web technologies, it is good practice to re-use IRIs (Internationalized Resource Identifiers) from existing knowledge bases, when possible [3].

In order to import existing IRIs into our ontology and our tool, we investigated three of the most popular knowledge bases: Wikidata, DBpedia [8], and

[3] `https://wikidata.org`

YAGO [10]. A recent study has compared the quality of these knowledge bases[4] according to various metrics [6]. The results of this study highlight that Wikidata is the top-rated knowledge base on the average of the considered metrics, with especially high scores on trustworthiness, consistency, timeliness, relevancy, and licensing. On the basis of these results and of an analysis we performed, we chose Wikidata as reference knowledge base, for the following reasons:

- it contains the largest number of entities (currently more than 50 million) when compared to the other knowledge bases;
- it is compatible with Semantic Web technologies such as RDF(S), OWL, and SPARQL [5];
- it is fully multilingual, with more than 39% of the entities having labels in multiple languages;
- it aims to collect not just statements about entities, but also the primary sources behind those statements (referenced statements are currently more than 75% of the total);
- it is fully integrated into Wikipedia and other Wikimedia projects, such as Wikimedia Commons, from which we can easily import non-structured data such as text and images;
- it adopts a Creative Commons Zero[5] license, equivalent to the public domain, making it easier to re-use the knowledge contained in it.

Notice that Wikidata also has a significant potential downside when compared to the other knowledge bases, i.e. its open and collaborative nature, similar to that of Wikipedia. The users of Wikidata can freely add and edit knowledge, including the class hierarchy, thus altering the ontology in unpredictable ways. However, both humans and bots frequently check the ontology for errors that can be due to users' mistakes or deliberate acts of "vandalism", and correct them.

3 The Wikidata Event Graph

Ideally, a suitable knowledge base for our purposes should contain: (i) historical entities such as people, places, and physical objects (ii) historical events connecting those entities. However, in most existing knowledge bases, including Wikidata, events are often represented in an implicit way. For instance, the knowledge base currently contains 4.52 million entities about people, but only 6,360 events of type "death", because most people's deaths are expressed implicitly through properties such as *P570 date of death*.

Many historical events, such as World War II, are represented explicitly in Wikidata, but they make up just 3% of the total number of entities. This is still a significant number overall (1.92 million)[6], but it is not enough for our

[4] We did not consider the other two knowledge bases analysed in the study (Freebase and OpenCyc), because they were both recently discontinued.

[5] `https://creativecommons.org/publicdomain/zero/1.0/`

[6] `https://bit.do/ewP9w`

purposes because in our ontology *all* events are represented explicitly. Indeed, in the four narratives[7] that were developed using our tool, the percentage of events that could be directly matched to Wikidata was less than 2%. This is significantly less than the percentage of related entities that could be found in Wikidata (69%).

In our view, the solution to this problem is the generation of what we call the Wikidata Event Graph (WEG), i.e. the Wikidata graph augmented with an explicit representation of all events implicitly expressed in it. Generating this graph would allow us to reference the events from our ontology and import them into our tool, offering the user a much more complete coverage of historical events.

In order to extract the WEG, we analysed all Wikidata properties[8] and compiled a list of the ones that, in our opinion, express implicit events. For instance, the property *P570 date of death* expresses an event of type "death".

We developed and implemented an algorithm that allows recognizing Wikidata properties that do not express events. From the current total of 5,234 properties, the algorithm removed several properties based on the following criteria:

1. the type of the property, i.e. meta-properties, properties that connect entities to other Wikimedia projects, obsoleted and deprecated properties, properties classified as "Wikidata property for an identifier", which simply link a Wikidata entity to an identifier in another knowledge base (for instance, the property *P214 VIAF ID* connecting an individual to its representation in the Virtual International Authority File[9]);
2. the datatype of the property's range literal, i.e. Web pages (datatype URL), numerical quantities (datatype Quantity), geographical features (datatypes GeoShape and GlobeCoordinate), media (datatype CommonsMedia), external IDs (datatype ExternalId), tabular data (datatype TabularData), and Wikidata properties (datatype WikibaseProperty).

At the end of this process, we obtained a list of about 1,000 candidate properties that could potentially express implicit events. Most of these (265) were applied to works, 158 were applied to people, 119 were applied to organisations, and the remaining ones were applied to other types of entities. In order to implement a case study, we decided to focus on the 158 properties about people. This would allow us to use as test case one of the narratives that had been previously constructed with our tool, i.e. the one about the life of Dante Alighieri[10].

[7] `https://dlnarratives.eu/narratives.html`

[8] The full list of properties is available at `https://www.wikidata.org/wiki/Wikidata:List_of_properties`. Another way to explore the properties is the Wikidata Property Explorer, available at `https://tools.wmflabs.org/prop-explorer/`.

[9] `https://viaf.org`

[10] `https://dlnarratives.eu/timeline/dante.html`

Table 1. Wikidata properties expressing implicit events about people's lives.

Event Type	Property ID	Property Name	Number of Events
Baptism	P1290	godparent	1,840
	P1636	date of baptism	
Birth	P19	place of birth	4,519,957
	P22	father	
	P25	mother	
	P40	child	
	P569	date of birth	
Creation	P50	author	546,048
	P57	director	
	P58	screenwriter	
	P61	discoverer or inventor	
	P84	architect	
	P86	composer	
	P87	librettist	
	P110	illustrator	
	P161	cast member	
	P170	creator	
	P178	developer	
	P800	notable work	
Death	P20	place of death	1,651,865
	P157	killed by	
	P509	cause of death	
	P570	date of death	
	P1196	manner of death	
Education	P69	educated at	763,823
	P184	doctoral advisor	
	P185	doctoral student	
	P512	academic degree	
	P802	student	
	P812	academic major	
	P1066	student of	
Election	P726	candidate	25,775
	P991	successful candidate	
	P3602	candidacy in election	
Foundation	P112	founder	22,359
Marriage	P26	spouse	94,726
Membership	P54	member of sports team	712,299
	P102	member of political party	
	P463	member of	
Occupation	P6	head of government	3,310,621
	P35	head of state	
	P39	position held	
	P106	occupation	
	P108	employer	
	P210	party chief representative	
	P286	head coach	
	P803	professorship	
	P1075	rector	
Residence	P263	official residence	60,372
	P551	residence	

4 An Initial Experiment

To perform our initial experiment, we ordered the list of 158 Wikidata properties obtained in the previous step by usage in the knowledge base. Among the most used, we selected the first 50 that clearly expressed events about a person's life. The list of 50 properties we considered is reported in Table 1.

We developed a software that, through the Wikidata Query Service[11], automatically extracts all events that were expressed implicitly by the properties of Table 1 from the Wikidata graph. As expected, the number of birth events is the most numerous (4.52 million), since every person has a birth. The second most numerous type of event is occupation (3.31 million), followed by death (1.65 million). The total number of events contained in the subset of the WEG that we generated is 11.71 million, thereby increasing the number of Wikidata events that can be linked from our tool by more than 600%.

The software removes duplicate events, i.e. identical events that can be extracted more than once through multiple properties, by applying the following criteria implemented using rules:

1. if two entities are linked by a direct property and also by its inverse, e.g. the properties *P802 student* and *P1066 student of*, the two events extracted from the properties are merged;
2. if two entities are linked by a symmetric property, e.g. *P26 spouse*, in both directions, the two resulting events are merged;
3. when two properties express the same event, e.g. if the property *P569 date of birth* and the property *P19 place of birth* are applied to the same person, the two resulting events are merged.

In the first version[12] of the narrative of the life of Dante Alighieri constructed using our tool, only one of 53 events (the Battle of Campaldino) was present in Wikidata as instance of the class *Q1190554 Occurrence*. After generating the WEG, we identified in it 31 more events that were present in the narrative. Therefore, the percentage of events in the narrative that could be automatically detected in Wikidata has increased from 1.9% (1 event) to 60.4% (32 events).

Major events such as the birth of Dante, the writing of the *Divine Comedy*, and the election of Pope Boniface VIII are all contained in the WEG, despite not being explicitly present in Wikidata as instances of the class *Q1190554 Occurrence*. Furthermore, the WEG contains 99 more events about Dante's life that are not present in the narrative built using our tool. Many of these are minor events, e.g. the writing of a sonnet, but it can still be useful to propose them to the user during the narrative building process.

We consider these results very promising, and anticipate that the coverage can be improved further by including more properties in our study. Furthermore, as more knowledge is added to Wikidata every day, the number of events in the WEG will increase.

[11] `https://query.wikidata.org`

[12] `https://dlnarratives.eu/timeline/dante.html`

5 Conclusions and Future Works

In this paper we have presented an initial study on the population of formal narratives through the import of events from the Wikidata knowledge base. We have analysed all Wikidata properties and used a subset of them to generate the Wikidata Event Graph (WEG), i.e. the graph of events that are expressed in an implicit way in the knowledge base.

As case study, we have taken into account one of the narratives that were built using our Narrative Building and Visualising Tool (NBVT), i.e. the life of Florentine poet Dante Alighieri. For this reason, we have focused on the detection of events about people's lives from Wikidata. Furthermore, we have generated a subset of the WEG containing 11.71 million events related to people's lives, thereby increasing the number of Wikidata events that can be linked from our tool by more than 600%. From this subset of the WEG we have extracted a subgraph of events related to the life of Dante, allowing us to increase the number of events that could be automatically detected in Wikidata from 1.9% (1 event) to 60.4% (32 events).

As future work, we plan to take into account other Wikidata properties related not only to people's lives but also to other topics of narratives such as scientific experiments and historical events, and generate a larger subset of the WEG from these properties. In order to perform a preliminary evaluation of our approach, we are also working to automatically extract events from Wikidata related to the other narratives that have been constructed using our tool.

Another issue that we aim to study is how to automatically identify events related to the subject of the narrative and propose them to the users for import in their narratives. In addition, we are currently investigating automatic narrative extraction from text. We believe that the WEG will prove very useful in this context, in particular to increase the recall of entity linking algorithms [9] applied to narrative texts.

References

1. Bartalesi, V.: An Ontology for Narratives. Ph.D. thesis, University of Pisa (2017)
2. Bartalesi, V., Meghini, C., Metilli, D.: A conceptualisation of narratives and its expression in the CRM. International Journal of Metadata, Semantics and Ontologies **12**(1), 35–46 (2017)
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked data: The story so far. In: Semantic services, interoperability and web applications: emerging concepts, pp. 205–227. IGI Global (2011)
4. Doerr, M.: The CIDOC Conceptual Reference Module: an ontological approach to semantic interoperability of metadata. AI magazine **24**(3), 75 (2003)
5. Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandečić, D.: Introducing Wikidata to the linked data web. In: International Semantic Web Conference. pp. 50–65. Springer (2014)
6. Färber, M., Bartscherer, F., Menne, C., Rettinger, A.: Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. Semantic Web **9**(1), 1–53 (2017)

7. Ferrara, A., Nikolov, A., Scharffe, F.: Data linking for the Semantic Web. International Journal on Semantic Web and Information Systems (IJSWIS) **7**(3), 46–76 (2011)
8. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., et al.: Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web **6**(2), 167–195 (2015)
9. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Transactions on Knowledge and Data Engineering **27**(2), 443–460 (2015)
10. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of the 16th international conference on World Wide Web. pp. 697–706. ACM (2007)
11. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)