Using Wikibase for Managing Cultural Heritage Linked Open Data Based on CIDOC CRM

Joonas Kesäniemi[1(✉)], Mikko Koho[1,2], and Eero Hyvönen[1,2]

[1] Semantic Computing Research Group (SeCo), Aalto University, Espoo, Finland
joonas.Kesaniemi@aalto.fi

[2] HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Helsinki, Finland

Abstract. This paper addresses the problem of maintaining CIDOC CRM-based knowledge graph (KG) by non-expert users. We present a practical method using Wikibase and specific data input conventions for creating and editing linked data that can be exported as CIDOC CRM compliant RDF. Wikibase is a proven and maintained software for generic KG maintenance with a fixed but flexible data model and easy-to-use user interface. It runs the collaboratively edited Wikidata KG, as well as increasing amount of domain specific services. The proposed solution introduces a set of data input conventions for Wikibase that can be used to generate CIDOC CRM compliant RDF without programming. The process relies on the aforementioned data input rules combined with generic mapping implementations and metadata stored as part of the KG. We argue that this convention over coding makes the system more easily approachable and maintainable for users that want to adhere to the CIDOC CRM principles, but are not ontology experts. As part of the preliminary evaluation of the proposed solution, an example on managing Cultural Heritage data in the military history domain with discussion on the limitations of the approach is presented.

Keywords: Wikibase · CIDOC CRM · Knowledge graph · Cultural heritage · Linked data

1 Introduction

While thousands of Linked Data datasets of Cultural Heritage (CH) have become openly available in the Linked Open Data (LOD) cloud[1] and elsewhere, an ever-more serious challenge is how to manage the knowledge graphs (KG) when the underlying ontologies and metadata evolve over time [4]. This is especially true for datasets that are products of unique research projects with minimal resources available for maintenance in the long run. When the graph is created using data

[1] https://lod-cloud.net/.

S. Chiusano et al. (Eds.): ADBIS 2022, CCIS 1652, pp. 542–549, 2022.
https://doi.org/10.1007/978-3-031-15743-1_49

exported from existing systems or from manually created sets of heterogeneous source files, it might be easy to make changes to the source data, but the transformation pipelines still require maintenance in order keep the graph version of the data relevant. In order to minimize the need for maintaining transformations, one can also choose to maintain the data graph directly. This is the approach taken, for example, in [9] in the context of culinary tradition and in [7] in archaeology domain. Both examples are using the CIDOC Conceptual Reference Model (CRM)[2] as their ontological foundation. CRM is a standard for interoperable information in the CH field [2]. It is an event-based top-level ontology with high expressive power to address the intricacies of describing CH data. CRM is also a complex ontology and its implementations produce equally complex networks of data, which makes creating easy-to-use tooling a challenge. The CRM events can make references to places where the event occurred, to the people and physical objects involved, and to the time-span of the occurrence. As the information of individual CH objects and actors relating to them is split into multiple events and a large amount of references between entities, making changes to the data is not as straight-forward as with more simple document or object-based data models. For example, adding a previously unknown birth date for a person would involve either creating a new birth event linked to the person, or modifying a birth event linked to the person, if one exists already.

This paper presents ongoing work related to the maintenance of CRM-based linked data using the Wikibase software[3]. Wikibase is developed to serve as the platform for Wikidata[4], which is a massive open and free knowledge base maintained by both humans and machines and hosted by the Wikimedia Foundation. Wikidata currently contains structured information about more than 90 million things and maintain an edit history of more than 1.6 billion actions, so it has been proven to be able to handle massive datasets if necessary. Wikibase is available as an open-source software suite and comes out-of-box with generic and simple user interface, search functionality and support for complete change history. It is designed for open and collaborative management of KGs and has been adopted by libraries, archives, and research groups for their data management needs [1]. We present a practical framework called WB-CIDOC for creating and editing CRM based data with Wikibase instance with an option to publish it as a custom CRM compliant RDF[5] dataset. Our approach tries to strike a balance between the structures required by the reference ontology and data input conventions needed to generate them by re-purposing features of the Wikibase data model. In Sect. 2 related works are first presented and the proposed method is introduced in Sect. 3. Experiences of using the method in our case study are then presented (Sect. 4). Finally, in conclusion, contributions of this paper and current limitations are summarized.

[2] https://www.cidoc-crm.org/.
[3] https://wikiba.se/.
[4] https://www.wikidata.org/.
[5] https://www.w3.org/RDF/.

2 Related Work

Existing works that combine technical solutions for manipulating KGs directly and use the CRM ontology are few, but there are some notable related works to be found. WissKI [10] is an open source Virtual Research Environment and content management system for Cultural Heritage data built on top of CIDOC CRM ontology. WissKI uses so-called ontology paths to map simple data inputs such as "date created" into more complete CRM data structures. Ontology paths are similar to the data input conventions presented in this paper. The main difference is how these shortcuts or mappings are used in the user interface. WissKI uses ontology paths to compile the input forms, whereas our approach relies on Wikibase's generic form interface. Also, in WissKI the ontology paths are added manually as opposed to a set of predefined input rules of WB-CIDOC.

Another more recent solution that combines KGs and CIDOC CRM is the ResearchSpace platform[6]. ResearchSpace advertises itself as a collaborative and contextualizing knowledge system that allows users to connect qualitative and quantitative research. It provides integrated tools for building dataset as well as semantic web applications. Many features of the ResearchSpace rely heavily on the CIDOC CRM ontology. It provides read and write access to the data through semantic forms, which in contrast to Wikibase's built-in generic implementation, must be implemented on a case-by-case basis. Examples [7] and [9] mentioned in the Sect. 1 are implemented with ResearchSpace. More detailed description of the system and more examples of research projects utilizing ResearchSpace to collect, enrich, and analyze data can be found in [8].

Gayo et al. [5] use Wikibase for representing and maintaining Shared Authority Files (SAF) in CIDOC-CRM in a project by Luxembourg Competency Network on Digital Cultural Heritage. SAFs are used by memory organizations to maintain authoritative records of identities of important entities, such as persons, places, and organizations. The approach is identical to ours, as it uses Wikibase qualifiers to provide data for linked resources. They even provide example SPARQL queries for mapping between the Wikibase data model and CIDOC CRM. However, these queries seems to be manually created examples, where as we use the metadata stored in the KG to automatically generate necessary transformations. The work also has a much narrower scope than our work, focusing solely of managing person related SAF data, which can be modelled with only a small subset of CIDOC CRM, namely the Appelations and their related Symbolic objects. In fact, their work is complementary to ours, since our WB-CIDOC approach does not consider CIDOC CRM Appelations yet.

3 WB-CIDOC – Wikibase for CIDOC CRM

This section describes the WB-CIDOC approach for using Wikibase as the source for CIDOC CRM-compliant RDF datasets. The potential user of the WB-CIDOC is someone who needs to be able to create CRM-based data and

[6] https://github.com/researchspace/researchspace.

might even understand why, but who does not need nor want to understand the ontological nuances of the CRM. Our potential user is also more inclined to implement data processing automations through graphical user interface instead of scripting.

WB-CIDOC consists of two parts: Wikibase items and properties that make up the schema for of the data to be inputted with metadata mapping to CRM and a set of simple data input conventions that take advantage of Wikibase's data model and user interface to denote shortcuts for resources required by the CRM. The former provides the model and mapping of the domain specific data to the CRM ontology and the latter gives the rules on how to use to the model in different situations. Finally, an external software component is used to implement the mapping from Wikibase's data model to the CRM RDF output with the help of the aforementioned schema and data adhering to the rules as described below. WB-CIDOC is a generic approach and should be suitable for any CRM project that can work with currently supported CRM features.

Stemming from its collaborative nature, Wikibase's data model is designed to represent statements about items of interest and their references. It also allows for every statement to be qualified with one or more additional statements, similarly to RDF reification [6]. For example, the Wikidata item for Douglas Adams[7] contains a statement with property `occupation` and value `novelist` with a qualifying information `start time 1979`, i.e., a qualifier provides additional information about its linked statement. A more detailed description of the Wikibase's data model can be found in [11].

Items and properties are the basic entity types in Wikibase. In order to be able to denote that some item "X" represents a person, one could, for example, first create an item for `Person` and a property `instance of`, and then add a statement `instance of Person` to the item "X". WB-CIDOC requires at least three properties to function: `instance of`, `URI mapping` and `event type`. The property `instance of` is used as explained above to link an item to a specific type or class of entities. Properties `event type` and `URI mapping` are used in the RDF export process to identify and generate CRM events, and to translate Wikibase generated URIs with Q and P numbers to external schemas, respectively. Other items for types and properties are determined by the domain model of the data to be inputted. See Sect. 4 for an example from the military history domain. When the required and domain specific properties and items are in place, it is possible to start inputting the actual data. It should be noted that the domain model can be modified at any point if and when the need arises.

Data input rules of the current alpha version of the WB-CIDOC are focused on creating CRM events and time-spans. We are reporting on ongoing work and more rules can be added in the future.

A so called *implicit event* is created by adding a statement to an item with a property that has an aforementioned event type statement. For example, we can have a property called `was born` with identifier P2 with following statements: `event type crm:E67_Birth` and `URI mapping crm:P98i_was_born`. The data

[7] https://www.wikidata.org/wiki/Q42.

type of the property `was born` must be `monolingual text`[8]. Finally, the implicit event denoting the birth for an item describing `John Doe` could be added with the following statement: `John Doe was born ''Birth of John Doe''`. The value of the `was born` property would become the preferred label of the generated `crm:E67_Birth` event and it would be linked to the person with the property `crm:P98i_was_born` as metadata recorded as part of the `was born` (P2) property. An implicit event can be further described by adding qualifiers to the statement. Continuing with the previous example, we might want to record the birth place and mother of John Doe. This can be achieved by adding qualifiers to the "was born" statement. Again `URI mapping` values of both properties `took place at` and `by mother` would be used as part of the RDF generation to map Wikibase URIs to something that is valid CRM. It should be noted here, that `crm:Event` instances are only generated for properties that are associated with event type metadata and they is used to describe items. Therefore, using the same property as a qualifier would not trigger an event generation.

If the event is linked to multiple actors or things, the user must decide on which related item to add the implicit event statement. For example, `E67_Birth` event involving a child and a mother, can be added either to the mother's (Q1) side as `Q1 gave birth ''Birth of Y''` with the qualifier `brought to life Q2` or to the item representing the child (Q2) as `Q2 has birth event ''Birth of Y''` with the qualifier `by mother Q1`. The problem with this approach is that the event is not directly visible from the item page of the other items involved. For example, in the latter case there would not be a link from mother to the child. These connections can be discovered on-demand by using Wikibase's "What links here" feature. Another option is to make the participation in an event explicit by adding inverse links to other event participants. For example from mother to child with a statement `brought to life John Doe`. This kind of property pairs are handled by the mapping implementation in a way that both persons refer to the same event even if both `was born` and `brought to life` properties contain `event type` metadata.

Another basic data input convention is used to handle time-spans. Time-span related properties, such as crm:P82a_begin_of_the_begin and `crm:P79_beginning_is_qualified_by`, can be added directly to implicit events as qualifiers. During export a new resource with the type `crm:E52 Time-Span` is created from all the properties from the qualifiers where their `URI mapping` metadata refers to a CRM property with the domain `crm:E52_Time-Span`.

4 Case WarSampo – Finnish World War II Data

WarSampo[9] [3] is a semantic portal and a linked data service about Finland in the Second World War. It consists of a harmonized KG in a SPARQL endpoint assembled from several heterogeneous sources, as well as a web portal with nine

[8] https://www.mediawiki.org/wiki/Wikibase/DataModel.

[9] Project: https://seco.cs.aalto.fi/projects/sotasampo/en; portal: https://www.sotasampo.fi/en.

application perspectives to the underlying KG. The WarSampo KG makes extensive use of CRM and its event-based model. The National Archives of Finland has received significant amount of correction suggestions to the data through the portal, and WB-CIDOC is one of the approaches we are currently experimenting with for incorporating changes back to the KG.

WB-CIDOC's implicit events are a good match for events that are not significant enough to warrant their own item and involve limited amount of participants. Let's take the aerial victory `<http://ldf.fi/warsa/events/event_lv32101>` as an example. The event has the pilot and his squadron as participants with details about the location, time period and aircraft involved. Figure 1 shows a partial description of the data in Wikibase using WB-CIDOC. In addition to the statements visible in the figure, the item `John Doe` is defined as `instance of Person` and `Squadron 32` as `instance of MilitaryUnit`. Also, `Person` and `MilitaryUnit` are defined with metadata `URI mapping crm:E21_Person` and `URI mapping crm:E74_Group` respectively.

Fig. 1. Partial description of the event `http://ldf.fi/warsa/events/event_lv3288` in Wikibase user interface.

The event described in Fig. 1 includes two participants which both use the property `participated in` that has event type `crm:E5_Event`. However, only one instance of an event is generated since mapping implementation can detect through qualifiers that they should refer to the same event. Figure 2 shows a snippet from the generated RDF export, which includes the processing of an implicit event with time-span.

5 Discussion

This paper presented an approach called WB-CIDOC for maintaining cultural heritage data that is compiliant with CIDOC CMR using the Wikibase platform. The presented approach simplifies data input task by introducing certain input conventions that can be used to infer some of the required CRM resources,

```
warsa:Q2 skos:prefLabel "John Doe" ;
    a crm:E21_Person ;
    crm:P11i_participated_in warsa-events:Q2-3690b3d7 .
warsa:Q3 skos:prefLabel "Squadron 32" ;
    a crm:E74_Group ;
    crm:P11i_participated_in warsa-events:Q2-3690b3d7 .
warsa-event:Q2-3690b3d7
    a crm:E5_Event ;
    skos:prefLabel "Aerial victory in Seiskari"@en ;
    crm:P11_had_participant warsa:Q3, warsa:Q2 ;
    crm:P14_carried_out_by warsa:Q3 ;
    crm:P7_took_place_at warsa:Q4 ;
    crm:P4_has_time-span warsa:Q2-3690b3d7-ts-1 ;
warsa-event:Q2-3690b3d7-ts-1
    a crm:E52_Time-Span ;
    crm:P82a_begin_of_the_begin "1942-04-03"^xsd:date ;
    crm:P82b_end_of_the_begin "1942-04-03"^xsd:date ;
```

Fig. 2. Subset of RDF output of the WarSampo example with a generated event instance and time-span.

hence lowering the amount of manually created and maintained entities. It also simplifies the mapping process, because the data input is based on Wikibase's generic user interface and transformation from internal data format to CRM RDF is done by generic implementations generated from the content of the KG.

WB-CIDOC currently focuses on basic CRM event descriptions and is not suitable for complex data in its current form. Some of the limitations stem from Wikibase's data model, such as support for only one reference per implicit event. For example, it is not possible to denote that time-span information came from "source X" and participant information from "source Y". The use of monolingual text as the data type of properties with event type metadata means that event labels are currently generated in only one language. One possible workaround could be to define special label property for adding other languages as qualifiers. Also, RDF export related event generation cannot currently properly link entities to shared events if there are more than one implicit event with same property and pair of participants. This kind of situation will lead to duplicate events instead of one shared event.

Implementation of the proposed solution is currently in an experimental stage, as we are working with the National Archives of Finland to develop different ways of maintaining the WarSampo data. Applying WB-CIDOC approach to other published datasets with more elaborate use of CRM would be an interesting area of further research as well as a step towards validation of the proposed solution.

Acknowledgements. Our work is funded by the Memory Foundation for the Fallen, National Archives of Finland, and FIN-CLARIAH project of the Academy of Finland. Computing resources of the CSC – IT Center for Science are used.

References

1. Diefenbach, D., Wilde, M.D., Alipio, S.: Wikibase as an infrastructure for knowledge graphs: the EU knowledge graph. In: Hotho, A., et al. (eds.) ISWC 2021. LNCS, vol. 12922, pp. 631–647. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-88361-4_37
2. Doerr, M.: The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. AI Mag. **24**(3), 75–75 (2003)
3. Hyvönen, E., et al.: WarSampo data service and semantic portal for publishing linked open data about the second world war history. In: Sack, H., Blomqvist, E., d'Aquin, M., Ghidini, C., Ponzetto, S.P., Lange, C. (eds.) ESWC 2016. LNCS, vol. 9678, pp. 758–773. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-34129-3_46
4. Koho, M., Ikkala, E., Heino, E., Hyvönen, E.: Maintaining a linked data cloud and data service for second world war history. In: Ioannides, M., et al. (eds.) EuroMed 2018. LNCS, vol. 11196, pp. 138–149. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01762-0_12
5. Gayo, L., Michelle, P.: Representing the luxembourg shared authority file based on cidoc-crm in wikibase (2021). https://www.youtube.com/watch?v=MDjyiYrOWJQ (Accessed 20 Jun 2022)
6. Manola, F., Miller, E., McBride, B., et al.: Rdf primer. W3C Recommendation **10**(1–107), 6 (2004)
7. Marlet, O., Francart, T., Markhoff, B., Rodier, X.: Openarchaeo for usable semantic interoperability. In: ODOCH 2019@ CAiSE 2019 (2019)
8. Oldman, D., Tanase, D.: Reshaping the knowledge graph by connecting researchers, data and practices in researchSpace. In: Vrandečić, D., et al. (eds.) ISWC 2018. LNCS, vol. 11137, pp. 325–340. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00668-6_20
9. Partarakis, N., et al.: Representation and presentation of culinary tradition as cultural heritage. Heritage **4**(2), 612–640 (2021). https://doi.org/10.3390/heritage4020036
10. Scholz, M., Goerz, G.: Wisski: a virtual research environment for cultural heritage. In: Raedt, L.D., et al. (eds.) ECAI 2012–20th European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS-2012) System Demonstrations Track, Montpellier, France, 27–31 August 2012. Frontiers in Artificial Intelligence and Applications, vol. 242, pp. 1017–1018. IOS Press(2012). https://doi.org/10.3233/978-1-61499-098-7-1017, https://doi.org/10.3233/978-1-61499-098-7-1017
11. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014)