Undefined 0 (0) 1
IOS Press

World War 1 as Linked Open Data

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Juha Törnroos [a,*], Eetu Mäkelä [a], Thea Lindquist [b], and Eero Hyvönen [a]
[a] *Semantic Computing Research Group (SeCo), Aalto University and University of Helsinki, Espoo, Finland*
[b] *University of Colorado Boulder, CO, USA*

Abstract. The WW1LOD dataset is primarily a reference dataset meant to bind together collections dealing with the First World War. For this purpose, the dataset gathers events, places and agents related to the war from various authoritative sources. These are then made available for indexing and other use through a variety of interfaces and APIs. Additional information on the entities is also collected, in order to be able to answer more complex questions relating to them. The approach is being evaluated using a concrete WW1 online collection.

Keywords: linked data, historical data, modeling, dataset

1. Introduction

The WW1LOD dataset is a reference dataset meant primarily to bind together historical collections dealing with the First World War. The dataset is additionally structured to provide a sensible common viewpoint into all such material linked, resulting in improved access to and context for these collections.

The origins of the dataset are in user needs research undertaken at the University of Colorado Boulder (CU) with the goal of improving the findability and usability of digitized collections of primary sources, i.e. documents contemporary to or reported by those who experienced historical events. To better understand the problems humanities researchers and students encounter in utilizing such collections, researchers at CU conducted 21 semi-structured interviews with faculty and students [8]. The major user needs identified were better support for: 1) locating documents and data that are relevant to a particular topic within distributed online collections; and 2) evaluating the items found by situating them in their cultural-historical context. In addition, problems were identified with crossing language barriers, as well as with ambiguities and variants in names, such as place names changing over time or multiple forms of a person's name being used in different documents.

As Linked Data seemed a promising technology to tackle these problems, the CU researchers contacted the Semantic Computing Research Group at Aalto University about collaborating on this research.

In thinking about ways to interlink online historical collections, the project chose to follow the path highlighted by the ISO standard CIDOC-CRM [4], which has become widely accepted by cultural heritage institutions as a basis for integrating sources. The core idea of CIDOC-CRM is to link the collection items to their real world context through events they participate in or reference. These events, and their participants, places and times then provide a contextual framework that binds the items together. This seemd to fit well to primary sources, as their value is presicely in how they view and relate to actual historical events and their participants.

However, CIDOC-CRM only asserts that common events, agents (e.g., people and organizations), places and times are important and tells how they can be encoded. To maximize interoperability and form a rich interlinked network, the identifiers of these entities still

*Corresponding author. E-mail: juha.tornroos@helsinki.fi

need to be shared. The real work is then in creating suitable reference vocabularies from which to source those identifiers.

To focus the work, a constraining subject domain had to be selected. As historical interest in the First World War is on the rise in anticipation of the upcoming centenary (2014-2018), and a digitized collection of primary-source documents was newly available at CU[1] [10], WW1 was selected as the domain of interest for a case study. Work then moved on to identifying suitable sources from which to draw the necessary common events, agents, places, times and other information pertinent to that domain.

2. Dataset Description

The main types of objects found in the WW1LOD dataset are described in Table 1 along with their instance counts, while the core data model is depicted in Figure 1. As stated, the core data model closely follows CIDOC-CRM, with events at the center linking together participating agents, places and times.

Table 1
Core classes in the dataset.

Type	Nr.	Example
Event	690	"Battle of the Aisne, 1914"
Event type	47	"Naval Operations"
Place	1380	"Aisne (River)"
Agent	530	"13th Cavalry Brigade"
Time	157	"02/01/1917 - 06/09/1918"
Keyword	29	"Prussian Militarism"
Theme	5	"Naval history"

Supporting these core classes are themes and keywords. Of these, the themes are used to categorize historical events into major classes, mostly for user interface purposes. The keywords exist to provide non-event thematic foci for linking. This need was discovered early on in indexing the primary sources. Often the sources would reference events in general, for example, talking about general wartime hunger and malnutrition, or the resistance of the Belgian Catholic Church to German occupation. It was thus often useful to index the documents as talking about event types such as "resistance", respectively. In these cases the links are less direct (by one level), but this approach still facilitates the discovery of documents that may relate to a particular instance of resistance. By the same token, the keyword "agriculture" may be used to link documents dealing with agriculture to events affecting it.

[1]The collection can be browsed online at `http://libcudl.colorado.edu/wwi/index.asp`

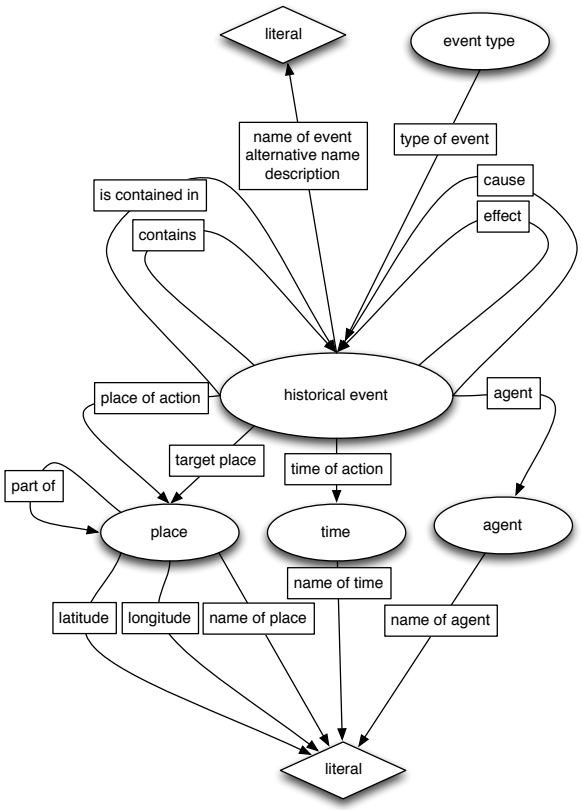

Fig. 1. The core data model of the dataset.

The actual instances included in the dataset come from multiple sources, as listed in Table 2. First, to provide a useful common base, general events, places and agents pertinent to the whole war were included.

An authoritative top-level framework of 326 wartime events was provided by the Imperial War Museum's First World War Centenary Partnership. The source staff most heavily relied upon to compile this event timeline was the official British series on the history of the war, the History of the Great War Based on Official Documents, particularly the volume Principal Events, 1914-1918 [**?**], with additional published works used to verify dates and facts. Unfortunately, these events were not linked to either places or agents however.

To overcome this limitation, a separate catalogue of 261 events selected by domain experts was built for richer description. They are drawn from vari-

Table 2
Contents of the dataset

Top-level events (326)	
Properties:	Name of event, Time of action, Description, Theme
Source:	Imperial War Museum (IWM) First World War Centenary Partnership
General events (261)	
Properties:	Name of event, Description, Agent, Time of action, Place of action, Is contained in, Contains, Cause, Effect
Source:	Imperial War Museum First World War event list, [1], CU Libraries
Atrocity events in Belgium (104)	
Properties:	Name of event, Agent, Time of action, Place of action, Combat related, Deportations, Human shields used, Panic, Destroyed buildings, Killings
Source:	[6]
German army structure (473)	
Properties:	Name of agent, Part of
Source:	[9], [2]
Geography of Belgium and France (1312)	
Properties:	Name of region, Part of
Source:	IWM Western Front geographic keywords, GeoNames, CU Libraries
Belgian statistical data for the war years (12)	
Properties:	Population of both genders in years 1914-1918s
Source:	[3]
Polygons of Belgian provinces (56)	
Properties:	Name of region, Part of, Polygon
Source:	HISSTAT (Universities of Ghent, Brussels, Louvain-la-Neuve and State Archives of Belgium)

ous sources, including approved terminologies from the Imperial War Museum and British Army's Battle Nomenclatures Committee as well as a custom term list on Belgium and WW1, derived in part from Patrick Lefevre [5]. These events are linked to the top level events where appropriate, resulting in 46 *owl:sameAs* links. In addition, these events have been linked by automatic means to DBPedia[2], with a little over 100 *owl:sameAs* relationships. These latter links are however still to be validated by domain experts. It should also be noted that in general, the material on Wikipedia is not sufficiently strictly quality-controlled to satisfy the needs of history researchers. This also explains why DBPedia in general has not been used as a main source for concepts and entities related to the war.

Regarding places of wartime events, a key point is that places are temporal instances, i.e. they can change over time. Current geographical datasets such as GeoNames are thus not directly applicable to a historical case like WW1, as the documents and events often refer to the place names used at the time, or in a particular language.

As a core source of contemporary place name data, locations related to the war were gathered from the Imperial War Museum's WW1 geographical keyword vocabularies. These vocabularies also contains a partonomy structure for the locations, which was brought into the dataset. For example, the dataset contains that Pommeroeul (village) *is part of* Hainaut (municipality) *is part of* Belgium (country). Places missing from the vocabularies were annotated by hand during the process.

Coordinate information for these places was sourced through two means. First, 1248 contemporary place names were automatically determined to directly match their modern equivalents in GeoNames based on identical names and hierarchy information. These sameAs-links were then evaluated by domain experts for accuracy. Second, wartime boundaries of Belgian provinces were obtained from HISSTAT, an inter-university network in Belgium.

The focus on Belgium apparent from the second source was brought on by the intuition that adequately responding to particular reasearch questions would require more detailed context than available in the general thesauri and registries. To demonstrate how such more specialized contexts can be fitted to the general frame and utilized, a particular subtopic was selected as a focus. In analyzing the document collection of CU, much material related to the topic of atrocities committed by the German armies in Belgium was discovered. Thus, sources giving real world information on that were particularly scoured when collecting material for the dataset.

Historical context and uncertainty also creates complexity for modeling times, as it is sometimes hard to say with absolute conviction when a certain event took place. For this reason we used a model for presenting time which supports a level of uncertainty in the encoding. In other words this means that it is possible to present a timestamp for e.g. "at the beginning of the year 1917" by specifying four temporal points: the earliest possible start time, latest possible start time, ear-

[2] http://www.dbpedia.org/

liest possible end time and latest possible end time. By using such timeframes, analyses and visualizations of the temporal relationships between war events do not miss events with uncertain dates.

Returning to the focus area of German operations in Belgium, sources were scoured for details on German army structure to fill in the necessary agent information for the use case. Domain experts did extensive background work to ensure that the information about these military units is valid and in line with existing research knowledge. Details on atrocity incidents, such as the number of killings or number of destroyed buildings were sourced from Horne and Kramer's tome [6] which is the standard on atrocities committed by German army units in Belgium in 1914. To understand if these events had any impact on the Belgian population, wartime population statistics for Belgian provinces was sourced from the Annuaire Statistique de la Belgique [3].

3. Dataset Access

The dataset is being regularly updated, enriched and maintained by CU researchers using the SAHA web-based collaborative metadata editor for RDF data [7]. Besides editing, SAHA also provides browsing, exporting and SPARQL-endpoint functionalities. For the WW1LOD project, these functionalities can be accessed at the following web addresses:

SAHA Browser & Editor View
`http://purl.org/ww1lod/saha`
HAKO Faceted Search
`http://purl.org/ww1lod/hako`
RDF Dumps
`http://purl.org/ww1lod/dumps`
SPARQL Endpoint
`http://purl.org/ww1lod/sparql`

Additionally, the dataset has been loaded onto the ONKI ontology server [11], which provides vocabulary to participating indexing systems through its APIs as well as browsing functionalities of its own at `http://onki.fi/onkiskos/ww1/?l=en`.

Individual instances in the dataset are also defined in the `http://purl.org/ww1lod/` namespace using unambiguous computer-generated identifiers (e.g. `http://purl.org/ww1lod/event/106` ffor the Battle of Albert in 1916) and are redirected using the purl.org service to their corresponding views in SAHA. The namespace for all schema properties is `http://purl.org/ww1lod/schema#`.

The dataset is published under a CC-BY-SA Creative Commons license [3], which allows sharing and remixing of the dataset with attribution to the original licensors. We wish that all organizations mentioned in Table 2 are referenced in case of sharing the dataset. The license also allows altering or redistributing the dataset with the same or similar license.

4. Usage Examples

As stated, the primary purpose of the dataset is as a reference vocabulary to which institutions can link their collections dealing with the First World War. Here, the concept selection functionalities provided by the ONKI ontology library service at `http://www.onki.fi/` are key. ONKI allows participating institutions to integrate concept picking functionality from the vocabulary to their indexing system either through SOAP or REST web services, or through a Javascript widget.

As for demonstrating the benefits gained from such linking on the CU's WWI Collection Online, there are still some unaddressed issues that prevent the dataset from being exploited to its full potential, despite the project already containing all the necessary entities for rich linking between the documents.

Key among these is the fact that the CU collection is currently indexed only with event types and keywords. On the WW1LOD side on the other hand, only some agents and places are related to these, while the events themselves are not. Further, the IWM top-level events refer mainly only to countries as agents and not to the more detailed instances related to individual units, people, etc., in the dataset.

Despite these limitations, however, it is still possible to make interesting queries of the data. For example, the following SPARQL query orders units of the German 3rd Army by the number of atrocities in which they were involved in Belgium (e.g. showing a disproportionate number for the 101st to 106th Infantry Regiments):

```
PREFIX wwls: <http://purl.org/ww1lod/schema#>
PREFIX wwlr: <http://purl.org/ww1lod/region/>
PREFIX wwla: <http://purl.org/ww1lod/agent/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT ?unit_name (COUNT(?e) AS ?atrocities) {
 ?e a/rdfs:subClassOf* wwls:AtrocityEvent .
```

[3] `http://creativecommons.org/licenses/by-sa/2.0/`

```
  ?e wwls:place_of_action/wwls:regionPartOf* wwlr:58 .
  ?e wwls:agent ?a .
  ?a wwls:subOrganizationOf* wwla:55 .
  ?a skos:prefLabel ?unit_name .
}
GROUP BY ?unit_name
ORDER BY DESC(?atrocities)
```

The following query, on the other hand, shows all documents relating to themes also associated with the Prussian general Friedrich Adolf Julius von Bernhardi (e.g. "Prussian militarism at work", a letter written by Henry Cleary, the Roman Catcholic Bishop of Auckland, on the theme of "Prussian militarism"):

```
PREFIX wwls: <http://purl.org/ww1lod/schema#>
PREFIX wwla: <http://purl.org/ww1lod/agent/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT ?kl ?wl WHERE {
 GRAPH <http://purl.org/ww1lod/> {
  wwla:542 wwls:related ?kw .
  ?kw skos:prefLabel ?kl .
 }
 GRAPH <http://libcudl.colorado.edu/wwi/> {
  ?w wwls:related ?kw .
  ?w rdfs:label ?wl .
 }
}
```

5. Discussion and Future Work

The WW1LOD dataset currently contains a basic top-level framework of events, agents and places upon which an ever-richer reference dataset about the First World War is being built. In addition to this framework, the structure needed to add ever-deeper contextual knowledge about particular subtopics is in place. Thought has also been given to how collections can link not only to particular events, agents and places, but also to more general topics, be they unit types, keywords etc. All of these authoritative and quality-checked resources are available through the ONKI ontology library service for WW1 collection holders to index their collections with. As such, it can be said with confidence that the collection already represents the most comprehensive linked open reference related to the topic available.

However, it should be noted that the dataset is decidedly a work in progress. Just a little additional work in linking the dataset parts together and filling in missing information would allow more complex questions to be answered, as well as truly bring the links to the primary sources. As such, these will be the next tasks on the project agenda.

Another improvement that needs to be made is to align the currently separate schema with that of CIDOC-CRM proper. This could not be done before, as the official CIDOC RDF namespace was agreed upon only in November of 2012.

For the purpose of reaching out to institutions holding other WW1 collections, plans are in the works to build visually attractive interactive user interfaces on top of the collection to demonstrate its potential. Proper validation of our approach to linking together collections would also require the addition and annotation of at least one other relevant collection.

This raises the issue of the amount of work necessary to properly index a collection against such a rich vocabulary. To help address this challenge, we are looking into avenues for the automatic discovery of events and entities from collection sources. However, new problems surface in using these techniques, such as those caused by variant spellings and OCR errors. They might be mitigated, for example, by linking to person identifiers in the Virtual International Authority File (VIAF)[4] data collected from the records of various national libraries.

Acknowledgements

We would like to thank Michael Ortiz (CU) and Tuomas Palonen (Aalto University) for annotating the resources, Michael Dulock (CU) and Holley Long (CU) for their aid with the digital collection and metadata, and Martha Hanna (CU), Patrick Tally (CU), Alan Kramer (Trinity College Dublin), Sophie de Schaepdrijver (Pennsylvania State University) and Tammy Proctor (Wittenberg University) for their expert opinion on the content.

References

[1] Great Britain. Battles Nomenclature Committee. *Official names of the battles and other engagements fought by the military forces of the British Empire during the Great War, 1914-1919*. London, 1921.

[2] Hermann Cron. *Geschichte des Deutschen Heeres im Weltkriege 1914-1918*. Biblio-Verlag, Berlin, 1937.

[3] Belgium. Ministère de l'Intérieur et de l'Hygiène. *Annuaire statistique de la Belgique et du Congo Belge*, volume 46. Brussels, 1922.

[4] http://viaf.org/

[4] Martin Doerr. The CIDOC CRM – an ontological approach to semantic interoperability of metadata. *AI Magazine*, 24(3):75–92, 2003.

[5] Patrick Lefevre ed. *Belgique et la Première Guerre mondiale, bibliographie: (Brussels: Musée Royal de l'Armée, 1987-2001)*. Musée royal de l'Armée, 1987-2001.

[6] John Horne and Alan Kramer. *German atrocities 1914: a history of denial*. New Haven: Yale University Press, 2001.

[7] Jussi Kurki and Eero Hyvönen. Collaborative metadata editor integrated with ontology services and faceted portals. In *1st Workshop on Ontology Repositories and Editors for the Semantic Web (ORES 2010), the Extended Semantic Web Conference ESWC 2010*. CEUR Workshop Proceedings, Vol-596, 2010.

[8] Thea Lindquist and Holley Long. How can educational technology facilitate student engagement with online primary sources?: A user needs assessment. *Library Hi Tech*, 29(2):224–241, 2011.

[9] Georg Tessin. *Deutsche Verbände und Truppen*. Osnabrück, 1974.

[10] Juha Törnroos Eetu Mäkelä Thea Lindquist, Eero Hyvönen. Leveraging linked data to enhance subject access - a case study of the university of colorado boulder's world war i collection online. In *World Library and Information Congress: 78th IFLA General Conference and Assembly, Helsinki*. IFLA, http://conference.ifla.org/ifla78, August 2012.

[11] Kim Viljanen, Jouni Tuominen, and Eero Hyvönen. Ontology libraries for production use: The Finnish ontology library service ONKI. In *The Semantic Web: Research and Applications, 6th European Semantic Web Conference, ESWC 2009, Proceedings*, pages 781–795. Springer-Verlag, 2009.