World War 1 as Linked Open Data

Eetu Mäkelä, Juha Törnroos, Thea Lindquist, and Eero Hyvönen

Abstract. WW1LOD is a reference dataset, an application of the CIDOC CRM ontology, meant to link collections dealing with the First World War. For this purpose, the dataset includes events, places, agents, times, keywords and themes related to the war, based on over ten different authoritative data sources from providers such as the Imperial War Museum. The content is harmonized into RDF using event-based modeling and published as Linked Open Data service. Furthermore, the data is enriched by linking to other Linked Open Data sources, such as DBpedia and Geonames, as well as through manual and automatic annotation. Together, the harmonization and enrichments allow the dataset to facilitate more complex queries than could be supported through any of the original data sources alone. As a demonstration of use, the ontologies, data and online service have been used to create a contextual reader for researchers consulting primary-source documents in WW1 related collections. This system automatically links textual descriptions to contextual information based on Linked Data, for example, the phrase “sinking of the Lusitania” in a PDF document to related content in DBpedia/Wikipedia, WW1 Discovery, Europeana and the Digital Public Library of America.

Keywords: linked data, historical data, digital humanities, modeling, dataset, contextual reader

1. Introduction

Interest in the First World War (WW1) is high due to its centenary anniversary (2014–2018). In recent years, a spate of projects have been launched to publish related content on the web, such as Europeana Collections 1914–1918, 1914–1918 Online, WW1 Discovery, Out of the Trenches, CENDARI, and Muninn. The confluence of interest and content exposed as Linked Open Data in this historical domain provides an unprecedented opportunity to bind together and enrich the user experience with cultural heritage collections.

The WW1LOD dataset presented in this paper is an application of the CIDOC-CRM and other established ontologies to create a Linked Data reference dataset of historical places, people and events related to the First World War. The purpose of this dataset is to act as a controlled vocabulary or reference authority linking together disparate historical collections and data publications related to WW1. It is additionally structured to provide common points of entry into such collections, resulting in improved access to and context for them. As a case study, the dataset has been applied to a collection of WW1 primary-source documents.

In this article, the reasons for creating the dataset and major technological choices made are first presented. The content of the dataset and its accessibility are then described in sections 3 and 4 respectively. Next, related work and projects are discussed and application examples are given in sections 5 and 6. In conclusion, contributions of this work and directions for future research are summarized.

2. Reasons for Creating the Dataset

The impetus for creating a Linked Data reference dataset in the WW1 domain sprang from user needs research undertaken at the University of Colorado Boulder (CU) to improve the user experience in working...
with digitized collections of historical primary sources (that is, documents contemporary to or reported by those who experienced the historical events under study).

To better understand the problems humanities researchers encounter in utilizing these collections, researchers conducted 21 semi-structured interviews with CU faculty and students (Lindquist and Long, 2011). The major user needs identified were better support for: 1) locating documents and data relevant to a particular topic within distributed online collections; and 2) contextualizing the content, for instance, to gauge author bias or simply become familiar with the places, events and people mentioned in the documents. In addition, problems were identified with documents being in unfamiliar languages, as well as with ambiguities and variations in names, such as place names changing over time or a person being referenced by different names in different documents.

To respond to such needs, reference vocabularies are needed that go far beyond the broad and often minimal subject headings commonly found in library cataloguing. In thinking about ways to interlink online historical collections, project partners chose to follow guidelines set out in the ISO standard CIDOC-CRM (Doerr, 2003), which cultural heritage institutions have widely accepted as a means to integrate sources.

The core idea of CIDOC-CRM is to link collection items to their real-world context through the events they reference. These events, and the people, places and timeframes related to them then provide a contextual framework that links together related items. This approach seemed well-adapted to primary sources, as their historical value is precisely in their view of and relation to the historical events they describe.

However, CIDOC-CRM only indicates that common events, actors (for instance, people and organizations), places and timeframes are important and provides a framework for how they can be described. To achieve interoperability between different datasets, these entities’ identifiers still need to be shared. The real work, then, is in creating suitable reference vocabularies from which to source those identifiers, for example, for individual battles, historical places and army units involved.

Further benefits are derived if these reference vocabularies are also themselves richly interlinked, allowing for inferencing and navigation among the actors, places and events identified.

To test these ideas, a focused case study was necessary. CU’s newly available collection of WW1 primary-source documents, the WWI Collection Online, (Lindquist et al., 2013) was selected for concrete evaluation, in part due to the timeliness of interest in the WW1 domain.

In creating the reference dataset, much thought was given to incorporating authoritative, high-quality information sources. This decision stemmed from needs the expressed needs of intended users of the system. For example, while Wikipedia contains numerous articles related to military history, many historians do not regard it as a reliable resource, either for themselves or their students (Waters, 2007). Indeed, while Wikipedia seems reassuringly accurate compared to traditional reference sources in the field of science (Giles, 2005), there is evidence to support their skepticism when it comes to the historical domain. In a field in which scholarly accountability is a core tenet, historians tend to raise questions about Wikipedia editors’ lack of accountability, the unevenness of the quality of its articles, and a lack of references to academic sources in its historical articles in particular (Luyt and Tan, 2010; Rector, 2008).

WW1LOD project partners, therefore, sought to incorporate data from as many sources as possible that historians would regard as authoritative and quality-controlled. At the same time, in order to become a part of and integrate with the established Linked Open Data (LOD) cloud, a real attempt has been made to create equivalency relationships between the project’s data sources and relevant resources in the cloud.

3. Dataset Description

The main object types found in the WW1LOD dataset are described in Table 1 along with their instance counts, while the core data model is depicted in abstract form in Figure 1. As stated, the actual data is

[The collection is available at] http:// libcudl.colorado.edu/wwi/index.asp
encoded using the much more complex CIDOC-CRM RDFS scheme. In total, the dataset is comprised of 55,703 triples.

<table>
<thead>
<tr>
<th>Type</th>
<th>Nr.</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event</td>
<td>4755</td>
<td>“Battle of the Aisne, 1914”</td>
</tr>
<tr>
<td>Event type</td>
<td>36</td>
<td>“Naval Operations”</td>
</tr>
<tr>
<td>Place</td>
<td>1366</td>
<td>“Aisne (River)”</td>
</tr>
<tr>
<td>Actor</td>
<td>651</td>
<td>“13th Cavalry Brigade”</td>
</tr>
<tr>
<td>Time</td>
<td>1960</td>
<td>“02/01/1917 - 06/09/1918”</td>
</tr>
<tr>
<td>Keyword</td>
<td>25</td>
<td>“Prussian Militarism”</td>
</tr>
<tr>
<td>Theme</td>
<td>5</td>
<td>“Naval history”</td>
</tr>
</tbody>
</table>

Table 1: Core classes in the dataset

![Fig. 1. The dataset’s core data model](http://ldf.fi/ww1lod/)

The instances included in the dataset come from a variety of sources, as detailed in Table 2.

3.1. Events

As a first step, general events spanning the war years were included in WW1LOD to provide a useful common basis for linking among related datasets (Hyvönen et al., 2012). For this, an authoritative framework of 326 top-level wartime events was provided by the Imperial War Museum (IWM)’s First World War Centenary Partnership. The IWM is considered the major authority among cultural heritage institutions in the English-speaking world where the First World War is concerned. Thus both historians and cultural heritage professionals should consider the sourced vocabulary authoritative, and it is likely to be re-used by others who are preparing datasets in this domain.

8Available at [http://www.cidoc-crm.org/rdfs/cidoc_crm_v5.0.4_official_release.rdfs](http://www.cidoc-crm.org/rdfs/cidoc_crm_v5.0.4_official_release.rdfs). More detailed information on the schema and dataset content can be found on the WW1LOD dataset homepage at [http://ldf.fi/ww1lod/](http://ldf.fi/ww1lod/)

9This timeline was principally derived from 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* (Great Britain, Committee of Imperial Defence, 1922).
Table 2
Dataset content

**Top-level events** (326)

<table>
<thead>
<tr>
<th>Graph:</th>
<th><a href="http://ldf.fi/ww1lod/iwm/">http://ldf.fi/ww1lod/iwm/</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Properties:</td>
<td>Name of event, Time of action, Description, Theme</td>
</tr>
<tr>
<td>Source:</td>
<td>IWM First World War Centenary Partnership</td>
</tr>
</tbody>
</table>

**Rich events** (253)

<table>
<thead>
<tr>
<th>Graph:</th>
<th><a href="http://ldf.fi/ww1lod/main/">http://ldf.fi/ww1lod/main/</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Properties:</td>
<td>Name, Alternate names, Description, Agent, Time of action, Place of action, Is contained in, Contains, Cause, Effect</td>
</tr>
<tr>
<td>Source:</td>
<td>IWM First World War event list (Great Britain. Army Council. War Office. Battles Nomenclature Committee 1921), CU Libraries</td>
</tr>
</tbody>
</table>

**Atrocity events in Belgium** (101)

<table>
<thead>
<tr>
<th>Graph:</th>
<th><a href="http://ldf.fi/ww1lod/atr/">http://ldf.fi/ww1lod/atr/</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Properties:</td>
<td>Name, Agent, Time of action, Place of action, Combat related, Deportations, Human shields, Panic, Destroyed buildings, Killings</td>
</tr>
<tr>
<td>Source:</td>
<td>(Horne and Kramer 2001)</td>
</tr>
</tbody>
</table>

**Major World War I events** (573)

<table>
<thead>
<tr>
<th>Graph:</th>
<th><a href="http://ldf.fi/ww1lod/wikipedia/">http://ldf.fi/ww1lod/wikipedia/</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Properties:</td>
<td>Name, Time of action, Related Wikipedia page, Theatre of war</td>
</tr>
</tbody>
</table>

**Automatically extracted timeline of principal events of the war** (3604)

<table>
<thead>
<tr>
<th>Graph:</th>
<th><a href="http://ldf.fi/ww1lod/principalevents/">http://ldf.fi/ww1lod/principalevents/</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Properties:</td>
<td>Name, Time of action</td>
</tr>
<tr>
<td>Source:</td>
<td>(Great Britain. Committee of Imperial Defence 1922)</td>
</tr>
</tbody>
</table>

**German army structure** (473)

<table>
<thead>
<tr>
<th>Graph:</th>
<th><a href="http://ldf.fi/ww1lod/iga/">http://ldf.fi/ww1lod/iga/</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Properties:</td>
<td>Name, Unit type, Part of</td>
</tr>
<tr>
<td>Source:</td>
<td>(Tessin 1974, Cron 1937)</td>
</tr>
</tbody>
</table>

**Other actors** (181)

<table>
<thead>
<tr>
<th>Graph:</th>
<th><a href="http://ldf.fi/ww1lod/main/">http://ldf.fi/ww1lod/main/</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Properties:</td>
<td>Name, Alternate names, Organizational information, Relationship information</td>
</tr>
<tr>
<td>Source:</td>
<td>CU Libraries</td>
</tr>
</tbody>
</table>

**Geography of Belgium and France** (1312)

<table>
<thead>
<tr>
<th>Graph:</th>
<th><a href="http://ldf.fi/ww1lod/main/">http://ldf.fi/ww1lod/main/</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Properties:</td>
<td>Name of region, Part of, Coordinates</td>
</tr>
<tr>
<td>Source:</td>
<td>IWM Western Front geographical keywords, GeoNames, CU Libraries</td>
</tr>
</tbody>
</table>

**Belgian statistical data 1914–1918** (12)

<table>
<thead>
<tr>
<th>Graph:</th>
<th><a href="http://ldf.fi/ww1lod/bsta/">http://ldf.fi/ww1lod/bsta/</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Properties:</td>
<td>Population</td>
</tr>
<tr>
<td>Source:</td>
<td>(Belgium. Ministère de l’Intérieur et de l’Hygiène 1922)</td>
</tr>
</tbody>
</table>

**Polygons of Belgian provinces 1914–1918** (56)

<table>
<thead>
<tr>
<th>Graph:</th>
<th><a href="http://ldf.fi/ww1lod/bpol/">http://ldf.fi/ww1lod/bpol/</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Properties:</td>
<td>Name of region, Part of, Polygon</td>
</tr>
<tr>
<td>Source:</td>
<td>HISSTAT (Universities of Ghent, Brussels, and Louvain-la-Neuve, and State Archives of Belgium)</td>
</tr>
</tbody>
</table>

The IWM events, however, did not contain place or actor information. To overcome this limitation, a separate catalogue of some 250 domain expert-selected events was built for richer description, including the annotation of places, participating actors, and temporal relationships. The information is drawn from various sources, including approved terminologies from the IWM and the British Army’s Battle Nomenclatures Committee, as well as a custom term list on Belgium and WWI created expressly for this purpose.
These events were manually linked to the top-level events where appropriate, resulting in 46 owl:sameAs links. In addition, all events have been automatically linked to DBpedia (Lehmann et al., 2015), with a little over 150 owl:sameAs relationships. Domain experts validated the latter links using the Silk link discovery framework (Volz et al., 2009).

Finally, details on the “German atrocities” that took place in Belgium in 1914 were included. This information was based on a vast study by John Horne and Alan Kramer (Horne and Kramer, 2001), which is considered the standard work on the topic. Included data, for example, are the place and date of each incident, the involved army unit, the number of killings, and the number of destroyed buildings.

The focus on Belgium apparent in this source was based on awareness that general thesauri and registries alone cannot provide the depth resolution needed for detailed research. Thus, the project decided to experiment on how also more detailed sources could be integrated as part of the general framework. In analyzing CU’s document collection, a rich subset of material was identified related to the German occupation of Belgium, and particularly to the violence German soldiers perpetrated against Belgian civilians. This particular subtopic was therefore selected as a focus and related sources were particularly emphasized when building the dataset.

While the sources described above provide a quality-controlled core of important, richly described events, they do not cover the whole scope of the war. Thus two additional sources of events were included to increase the breadth of coverage. First, 573 war events were extracted from the Wikipedia time-line of World War I[11]. These provide an alternate viewpoint into major wartime events. Second, OCR and automatic extraction were applied to a list of events appearing in Great Britain. Committee of Imperial Defence[12], producing a total of 3604 events. Though these events were extracted from a high-quality source, they contain numerous erroneous entries due to OCR and extraction problems.

3.2. Actors

Actor information in the dataset is primarily derived from two sources. The first is information on the structure of the Imperial German Army, which comes primarily from Georg Tessin’s Deutsche Verbände und Truppe (Tessin, 1974). This work is recognized in the field as a standard work of reference on the topic. The German army data was particularly important as it allows links to be made between the units mentioned in the atrocity data and their organizational hierarchy for further contextualization and reasoning.

Additionally, CU domain specialists added actor information in conjunction with enriching the event network. During this work, they not only linked actors to events, but also to each other and the organizations to which they belong.

In expressing organizational and relationship information, in addition to CIDOC-CRM, the project makes use of multiple generally accepted vocabularies for such data: the W3C Organization Ontology[12], the RELATIONSHIP ontology[13], the FOAF schema[14], and the schema.org vocabulary[15].

3.3. Historical Places

Regarding the locations of wartime events, a key point is that places are temporal instances, that is, they can change over time. Current geographical datasets such as GeoNames are therefore not always directly applicable to a historical case like WWI, as the documents and events often refer to the place names...
used at the time. Additionally, the locations of battles often refer, for example, to geographical features such as ridges or rivers as opposed to the villages or administrative units of the same name that are more commonly found in modern datasets.

As a core source of contemporary place name data, war-related locations were gathered from the IWM’s WW1 Western Front geographical keyword vocabulary. This vocabulary also contains a partonomy structure for the locations, which was brought into the dataset. For example, the dataset contains that Pommereou (village) is part of Hainaut (municipality) is part of Belgium (country). Additional place instances were added manually during the event and actor enrichment 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 name and hierarchy information. Domain experts then evaluated these links, encoded as skos:closeMatch triples, for accuracy.

Second, wartime boundaries for the Belgian provinces were obtained from HISSTAT, a collaborative project of the Universities of Ghent, Brussels, and Louvain-la-Neuve, and the State Archives of Belgium. Their geographies are highly accurate, penetrating to the municipal level. This provincial coverage, in combination with information from GeoNames and other sources, allows items to be placed in their geographical context using boundaries accurately defined for the war years. To encode the border polygons, the GeoRSS vocabulary was used.

3.4. Time

Historical context and uncertainty also creates complexity for temporal modeling, as it is often difficult to state with absolute precision when a certain event took place. For example, actors on opposing sides of a battle tend to see the extent of that battle differently (Great Britain. Committee of Imperial Defence, [1922] explanatory notes), depending on if they include preparatory attacks leading up to the battle or the consolidation of gains following a successful attack in the timeframe.

For this reason, CIDOC-CRM’s temporal representation supports a level of uncertainty in encoding. In other words, this means that it is possible to present a timestamp, for instance, for "at the beginning of the year 1917" by specifying four temporal points: the earliest possible start time, the latest possible start time, the earliest possible end time, and the latest possible end time. By using such timeframes, analyses and visualizations of the temporal relationships between war events do not miss those with uncertain or ambiguous dates.

3.5. Keywords and Themes

Supporting the core classes – events, agents, places and timeframes – are themes and keywords. Themes categorize historical events into major classes, mostly for use in public-facing interfaces. Keywords, on the other hand, provide non-event thematic foci for linking, a need that was identified early on in the process of indexing the primary sources. Often the sources would refer to pervasive wartime events in general, instead of, or in addition to individual instances. Thus it was often useful to index the documents also as referencing wartime hunger and malnutrition as a whole, or the resistance of the Belgian Catholic Church to German occupation. In these cases the links are less direct (by one level), but still allow the discovery of related items of probable relevance. For instance, the keyword "agriculture" may be used to link documents dealing with agriculture to events affecting it.

3.6. Statistics

To round out the coverage on Belgium, population statistics for the Belgian provinces during the war years were sourced from annual figures published by the Belgian Interior Ministry (Belgium. Ministère de l’Intérieur et de l’Hygiène, 1922). To model these statistics, the project utilizes the W3C Data Cube vocabulary.17

3.7. Interlinking

While the project strove to include only information humanities scholars would consider authoritative, it was also desirable to link the dataset to the LOD cloud to promote connection and interoperability. To facilitate this outcome, events and actors were linked to their counterparts in DBpedia, as well as other Linked Data datasets covering WWI. In addition, places were associated with GeoNames, where appropriate. An overview of these links, which domain experts at CU manually verified for accuracy, is provided in Table 3.

Table 3

<table>
<thead>
<tr>
<th>Target</th>
<th>Nr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>IWM and other events (internal)</td>
<td>46</td>
</tr>
<tr>
<td>DBpedia events and actors</td>
<td>152</td>
</tr>
<tr>
<td>Out of the Trenches events</td>
<td>29</td>
</tr>
<tr>
<td>GeoNames places (skos:closeMatch)</td>
<td>1248</td>
</tr>
</tbody>
</table>

4. Dataset Access

The dataset is published via the Linked Data Finland (LDF.fi) data publication service (Hyvönen et al., 2014), which provides browsing, editing and visualization services on top of standard SPARQL and Linked Data browsing APIs. The main page for the project is at http://ldf.fi/ww1lod/ which describes both the API endpoints as well as the add-on service links in detail. Individual instances in the dataset are also defined in the http://ldf.fi/ww1lod/ namespace using unambiguous computer-generated identifiers (e.g., http://ldf.fi/ww1lod/a74d369d for Viscount Bryce) and support direct Linked Data browser access with content-type negotiation.

The dataset is available openly under a CC-BY-SA 4.0 Creative Commons license which allows sharing and remixing of the dataset with attribution to the original licensors. All organizations mentioned in Table 2 should be referenced in case of using the dataset. The license also allows the dataset to be altered or redistributed with the same or similar license.

The dataset is being regularly updated, enriched and maintained by CU researchers live on the endpoint using the SAHA web-based collaborative metadata editor for RDF data (Mäkelä and Hyvönen, 2014). For this reason, actual instance counts or even data model specifics may not match those cited in this paper, which represents a snapshot in time. The most current information, however, is always available from the dataset’s home page.

17 http://www.w3.org/TR/vocab-data-cube/
18 http://creativecommons.org/licenses/by-sa/4.0/
5. Related Work

As stated, the dataset’s primary purpose is as a reference vocabulary to which other projects, institutions and organizations can link their WW1 collections. To avoid duplicate work and promote uptake, the project has striven to interface with as many of the others in the field as possible, both those publishing reference vocabularies and those publishing primary data.

Europeana 1914–1918 has created a simple vocabulary for WW1 that includes a subset of the Library of Congress Subject Headings (LCSH) along with 81 additional concepts. This vocabulary, however, is still preliminary, as the added concepts occur in a localhost namespace, and currently do not link to one another. Furthermore, the thesaurus (as the LCSH) is quite general and does not refer to individual units or events but rather to general keyword concepts such as "War Crimes" or "Eastern Front Campaigns". Finally, the vocabulary has incorporated LCSH’s pre-Linked Data structure, which joins discrete aspects of information under a single identifier, such as having separate concepts named "World War, 1914–1918 – Social aspects – Great Britain" and "World War, 1914–1918 – Social aspects – Germany". Conversations are ongoing with Europeana about how the WW1LOD dataset could be aligned with their work.

The Trenches to Triples project has published Linked Data vocabularies of WW1 events, places and actors. In practice, however, these vocabularies do not parse as valid RDF, and even corrected, the data model used is not consistent. For example, the same concepts are sometimes referred to using literals, and at other times using object references. Furthermore, the event, actor and place lists are separate and do not contain cross-references, literal or otherwise. Nevertheless, the data has been corrected, parsed and loaded into a staging area at http://ldf.fi/ww1/ for further study and possible integration into WW1LOD.

The Muninn project has published information particularly pertaining to Canada in the First World War. As of this writing, however, the primary data dump available consists of a SPARQL result set in XML format listing all quads rather than the N-Quads format claimed on its page. As the WW1LOD project has not yet had time to implement the custom parsing necessary to turn this result set back into triples, the dataset has also not yet been fully evaluated. Aside from this primary data however, Muninn does maintain a number of well-thought-out ontologies relevant to the domain. These ontologies were consulted in deciding upon the data models to be used for WW1LOD, but in the end the decision was made to use the CIDOC-CRM and related ontologies as a better supported albeit more general option.

A second Canadian project of interest is Out of the Trenches, which in addition to publishing primary collection data also includes a reference vocabulary containing 64 richly described events and 324 actors relevant to the collection. While the actors are almost exclusively Canadian and did not appear in the WW1LOD data, 29 event equivalencies were discovered between the datasets and recorded in the WW1LOD project.

The CENDARI project aims to provide WW1 historians with tools to contextualize, customize and share their research. The project is only starting to ramp up its vocabulary integration work, but has expressed an interest in using the WW1LOD vocabulary for this purpose.

Finally, Historypin has expressed an interest in the WW1LOD data. Their intention is to use the project’s event timeline to highlight events that occurred one hundred years ago as part of its centennial coverage.

6. Examples of Use

To help illuminate the dataset’s content and structure, several sample SPARQL queries are posted on the dataset home page as well as given in Appendix A. One uncovers that the German 103rd, 104th and
106th infantry regiments were disproportionately involved in atrocity incidents compared to other units of the German 3rd Army. Another shows no clear relationship between population change at the provincial level in Belgium during the war years, either as relates to the number of atrocity incidents or to the total number of general wartime events. A final sample query shows all documents in CU’s WWI Collection Online relating to themes associated with the Prussian general Friedrich Adolf Julius von Bernhardi (e.g., “Prussian militarism at work”, a letter written by Henry Cleary, the Roman Catholic Bishop of Auckland). Such queries are not meant to give definitive answers, but rather to highlight patterns in the data that researchers may wish to investigate more deeply.

In demonstrating the broader usefulness of the WW1LOD dataset, a first problem is that only the CU WWI Collection Online has been directly linked to it. The equivalency links to other vocabularies offer little help, due to either a low level of overlap or a high level of generality in these vocabularies. As an example, a query for items related to events in West Flanders does bring in content from multiple sources—Europeana, the CU WWI Collection Online, and Out of the Trenches. However, this example had to be carefully selected. In reality, the Out of the Trenches data contains just a single individual event (the Battles of Ypres, 1917), while the Europeana vocabulary contains only three. In both collections on the other hand, there are a multitude of items merely referencing the war as a whole.

While discussions are ongoing to facilitate other institutions’ direct use of the WW1LOD dataset, its utility can be demonstrated through the project’s contextual reader prototype. This demonstrator uses various methods to overcome issues arising from the lack of detailed annotations in related collections. Visualized in Figure 2, the application is able to dynamically process HTML and PDF documents and search for matching entities in multiple vocabularies relevant to the WWI domain, such as WW1LOD, Out of the Trenches, a subset of DBpedia, and the Europeana 1914–1918 thesaurus. These entities are then highlighted for the researcher’s inspection. When they are moused over, a short description and an image or a map appear, as illustrated on the left-hand side of Figure 2a with the example of Captain Charles Algernon Fryatt. Clicking on the entity opens a side pane in the interface containing both added

---

Fig. 2. The contextual reader interface

---

Note 25: http://j.mp/1mSSOU
Note 26: http://j.mp/1mSSnZO
Note 27: http://j.mp/1mSVKQx
Note 28: http://j.mp/1EP1P5i
Note 29: Available at http://demo.seco.tkk.fi/ww1/
context and linked resources from multiple sources, such as WW1 Discovery, Europeana, the Digital Public Library of America (DPLA), and the European Digital Library.

The contextual reader uses two primary methods to address the problem of lack of detailed metadata and, in turn, of linked content. First, it uses dynamic entity extraction to provide context for source collections other than the CU WWI Collection Online, for instance, articles in 1914–1918 Online, an authoritative, freely available encyclopedia of WWI. Thus the WW1LOD dataset can be utilized to for example show how an event mentioned in one of these articles fits in the context of the war as a whole. For example, in Figure 2a, a time-line in the reader interface locates the shooting of nurse Edith Cavell in the context of 1) the IWM’s important top-level wartime events, 2) all events happening in the same timeframe, and 3) other wartime events that happened nearby. All of these events are also presented on a map.

The second way the contextual reader bypasses this problem is by extracting relevant textual search terms from the WW1LOD dataset for an entity of interest and then using the associated collections’ existing keyword search functionality to discover related content. This is the primary way the reader currently brings in content from sources often lacking detailed metadata like WW1 Discovery, Europeana, DPLA and the European Digital Library. Among the extracted data used in the query are multi-lingual labels for a place or event as well as other closely related entities such as actors in an event or other events that took place in the same location. In Figure 2a, this functionality can be seen on the right-hand side of the user interface. Metadata brings in a relevant primary source from the CU WWI Collection Online. Images of the burial of Captain Fryatt from Europeana are also found not through controlled vocabularies used in the metadata, but rather from a keyword match on his name that appears in a textual description of the images.

In preliminary tests with the reader application, the WW1LOD vocabulary performs significantly better than the other thesauri used, both in contextualizing the sources and in bringing in relevant supplementary content. The reasons for this outcome seem to be twofold. First, the vocabulary is more focused and rich in its field, therefore resulting in significantly higher recall and precision in locating concepts. Second, WW1LOD has made a conscious attempt to include alternative names and spellings for concepts, further increasing recall both in the sources themselves and the linked resources.

7. Discussion and Future Work

The WW1LOD dataset currently contains a relatively rich, quality-controlled framework of events, agents and places related to the First World War. Thought has been given to how collections can link not only to the entities contained in the dataset, but also to more general categories, be they unit types, keywords, etc. As such, WW1LOD currently represents the most comprehensive linked open reference available related to WW1.

Discussions are ongoing with various projects, notably Out of the Trenches, Europeana 1914–1918, DPLA, and 1914–1918 Online on how the WW1LOD dataset could be better integrated with their collections and vocabularies. Here, the sparse annotations and general keywords in existing collections present a continuing challenge to integration. Where the CU WWI Collection Online is concerned, this gap was bridged both by manual annotation and by utilizing entity extraction tools underlying the contextual reader (Mäkelä, 2014) to create structured metadata for entities mentioned in the source texts. A similar approach is likely feasible for other mainly textual collections.

The WW1LOD dataset is itself still a work-in-progress. While the project strives to include diverse material in order to demonstrate the potential of its approach, coverage for all entity types except events is sporadic. Further work is needed to include geographies for places other than Belgium, the structures of armies other than the German, and so on. The hope is that researchers who could benefit from inclusion of this information will add it, thereby enriching the dataset.

Future steps the project could take to facilitate this outcome is to reach out to more institutions holding WW1 collections. Further, the contextual reader interface and other means of access would require further development in order to demonstrate the usefulness of the approach to potential partners.
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 Schaeppdrijver (Pennsylvania State University) and Tammy Proctor (Wittenberg University) for their expert opinion on the content.

References


Appendix

A. Sample SPARQL queries

A.1. Units of the German 3rd Army engaged in atrocities in Belgium
A.2. Population change in Belgium during the war
GROUP BY ?place
{
?place skos:prefLabel ?placeName .
}
ORDER BY ?populationDifference

A.3. Works in the CU WW1 Collection related to themes that are also related to General von Bernhardi

PREFIX wwl: <http://ldf.fi/ww1lod/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX bf: <http://bibframe.org/vocab/>
SELECT ?keyword ?work WHERE {
SERVICE <http://ldf.fi/ww1lod/sparql> {
ww1:c951988d skos:related ?kw . # General Friedrich Adolf Julius von Bernhardi
}
?w bf:title ?work .
}

A.4. Material from three collections related to events that took place in West Flanders

PREFIX wwl: <http://ldf.fi/ww1lod/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdag3: <http://rdvocab.info/ElementsGr3/> PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX crm: <http://www.cidoc-crm.org/cidoc-crm/>
PREFIX event: <http://purl.org/NET/c4dm/event.owl#>
PREFIX dct: <http://purl.org/dc/terms/> SELECT ?eventLabel ?object WHERE {
{
SELECT ?aPossiblySubEvent (SAMPLE(?anEventLabel) AS ?eventLabel) {
?place (owl:sameAs|owl:sameAs|crm:P89_falls_within)* wwl:82d810d4 . # West Flanders
?anEvent crm:P7_took_place_at ?place .
}
GROUP BY ?aPossiblySubEvent
}
{
BIND(STR(?anEventLabel) AS ?plainEventLabel)
SERVICE <http://europeana.ontotext.com/sparql> {
}
} UNION {
SERVICE <http://ldf.fi/colorado-ww1/sparql> {
}
} UNION {
}
?aPossiblySubEvent (owl:sameAs|owl:sameAs)* ?equivalentEvent .
SERVICE <http://ldf.fi/ww1/sparql> {
}
ORDER BY ?eventLabel