Noname manuscript No.
(will be inserted by the editor)

WW1LOD - An application of CIDOC-CRM to World War 1 Linked Data

Eetu Mäkelä · Juha Törnroos · Thea Lindquist · Eero Hyvönen

the date of receipt and acceptance should be inserted later

Abstract The CIDOC-CRM standard indicates that common events, actors, places and timeframes are important in linking together cultural material, and provides a framework for describing them. However, merely describing entities in this way in two datasets does not yet interlink them. To do that, the identities of instances still need to be either reconciled, or be based on a shared vocabulary.

The WW1LOD dataset presented in this paper was created to facilitate both of these approaches for 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, and published as a Linked Open Data service.

While generally basing on CIDOC-CRM, some modeling choices used also deviate from it where our experience dictated such. In the article, these deviations are discussed in the hope that they may serve as examples where CIDOC-CRM itself may warrant further examination.

As a demonstration of use, the dataset and online service have been used to create a contextual reader application that is able link together and pull in information related to WW1 from e.g. 1914–1918 Online, Wikipedia, WW1 Discovery, Europeana and the Digital Public Library of America.

E. Mäkelä · J. Törnroos · E. Hyvönen
Aalto University, Finland
E-mail: first.last@aalto.fi

T. Lindquist
University of Colorado Boulder, USA
E-mail: first.last@colorado.edu

Keywords applying CIDOC-CRM · linked data · modeling · historical data · dataset · data interlinking

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[1], 1914–1918 Online[2], WW1 Discovery[3], Out of the Trenches[4], CENDARI[5], and Muninn[6]. 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. It is one of only a handful of existing application descriptions of CIDOC [3, 2, 14] that can give insight into how the standard is able to cover material of different types.

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.

[1] http://www.europeana-collections-1914-1918.eu/
[2] http://www.1914-1918-online.net/
[3] http://ww1.discovery.ac.uk/
[4] http://www.canadiana.ca/en/pcdhn-lod
[5] http://www.cendari.eu/about-cendari
[6] http://blog.muninn-project.org/

In this article, the reasons for creating the dataset and major technological choices made are presented first. Section 3 then describes the contents of the dataset, while section 4 discusses how the ontologies selected were applied to model that data. Access to the dataset is outlined in section 5. Next, related work and projects are discussed and application examples are given in sections 6 and 7, respectively. In conclusion, the contributions made by 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 [15]. 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 [6], 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 among different datasets, these entities' identifiers still need to be shared. Pointedly, it doesn't do much good for integration to have two crm:E21_Persons in different data sets with the name "Robert Lansing", if they are not linked to a common strong identifier that disambiguates the First World War era politician from the later actor with the same name. 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[7] [16], was selected as the focus for evaluation, in part due to current interest in the WW1 domain.

In creating the reference dataset, a concerted effort was made to incorporate authoritative, high-quality information sources. This decision stemmed from the expressed needs of humanists as an important prospective user group for the system. For instance, though Wikipedia content is readily available and contains a wealth of information on military history, additional information sources were sought since most historians do not regard Wikipedia as a reliable source for either themselves or their students [26]. Although it has proven surprisingly accurate when compared to traditional reference sources in the field of science, existing evidence supports their skepticism about its reliability in the historical domain [7]. In a field in which scholarly accountability is a core tenet, historians often raise questions about Wikipedia editors' lack of accountability, the uneven quality of its articles, and the dearth of references to historical academic sources [17, 23].

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, an attempt has been made to create equivalency relationships between the project's data sources and relevant resources in the cloud.

[7] The collection is available at `http://cudl.colorado.edu/luna/servlet/UCBOULDERCB1~58~58`

3 Dataset Description

The main object types found in the WW1LOD dataset are described in Table 1 along with their instance counts. More detailed information on the schema and dataset content can be found on the WW1LOD dataset homepage at `http://ldf.fi/ww1lod/`. In total, the dataset is comprised of 63,807 triples.

Table 1 Core classes in the dataset

Type	Nr.	Example
crm:E5_Event	4764	"Battle of the Aisne, 1914"
ww1:EventType	36	"Naval Operations"
crm:E53_Place	1365	"Aisne (River)"
crm:E39_Actor	739	"13th Cavalry Brigade"
crm:E52_Time-Span	1830	"02/01/1917 - 06/09/1918"
org:Post	69	"Pope of the Roman Catholic Church"
ww1:Keyword	21	"Prussian Militarism"
ww1:Theme	5	"Naval history"

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 [11]. 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[8]. 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.

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 WW1 created expressly for this purpose[9].

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 [13], with a little over 150 *owl:sameAs* relationships. Domain experts validated the latter links using the Silk link discovery framework [25].

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 [9], 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.

When it became available, an additional set of 190 events and attendant encyclopedia articles were incorporated from the timeline created for 1914 – 1918 Online[10]. 1914 – 1918 Online is an international collaborative project to create an authoritative, freely available online encyclopedia of the First World War. Released in October 2014, it boasts an editorial board made up of approximately 90 distinguished historical experts from 22 countries with responsibility for assessing and managing the quality of the resource. To revisit the point about integration needing to happen on an instance level, it should be noted that while this source was in fact provided to us as CIDOC-CRM compatible RDF, it still needed exactly the same mapping work to be done on it in order to be useful.

While the sources described above provide a quality-controlled corpus of richly described major war events, the more nuanced picture that could be gleaned from the inclusion of lower-level events was still missing. Thus two additional datasets were included for users who opt to increase their breadth of coverage at the possible expense of data quality. The first is a set of 573

[8] This 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* [8]

[9] Derived in part from Patrick Lefevre's standard bibliography on this topic published by Belgium's Musée Royal d'Armée [12]

[10] `http://www.1914-1918-online.net/`

Table 2 Dataset content

Top-level events (326)	`http://ldf.fi/ww1lod/iwm/`
Classes:	crm:E5_Event
Properties:	skos:prefLabel, dc:description, crm:P4_has_time-span, crm:P130_shows_features_of (Theme)
Source:	IWM First World War Centenary Partnership
Rich events (253)	`http://ldf.fi/ww1lod/main/`
Classes:	crm:E5_Event
Properties:	skos:prefLabel, skos:altLabel, dc:description, crm:P4_has_time-span, crm:P7_took_place_at, crm:P10_falls_within, crm:P17_was_motivated_by, crm:P11_had_participant, crm:P14_carried_out_by
Source:	IWM First World War event list, CU Libraries
Atrocity events in Belgium (101)	`http://ldf.fi/ww1lod/atr/`
Classes:	ww1:AtrocityIncident (a subclass of crm:E5_Event)
Properties:	skos:prefLabel, crm:P14_carried_out_by, crm:P7_took_place_at, crm:P4_has_time-span, ww1:killings, ww1:humanShields, ww1:deportations, ww1:panic, ww1:combatRelated, ww1:destroyedBuildings
Source:	[9]
Event timeline from 1914 – 1918 Online (190)	`http://encyclopedia.1914-1918-online.net/lod/`
Classes:	crm:E5_Event
Properties:	skos:prefLabel, crm:P4_has_time-span, e1418:E1418_P_Related_Article, e1418:E1418_P_Related_Image
Source:	1914 – 1918 Online
Major World War I events (573)	`http://ldf.fi/ww1lod/wikipedia/`
Classes:	crm:E5_Event
Properties:	skos:prefLabel, crm:P4_has_time-span, crm:P130_shows_features_of (Theatre of war), foaf:page (Wiki page)
Source:	`http://en.wikipedia.org/wiki/Timeline_of_World_War_I`
Automatically extracted timeline of principal events of the war (3604)	`http://ldf.fi/ww1lod/principalevents/`
Classes:	crm:E5_Event
Properties:	skos:prefLabel, crm:P4_has_time-span
Source:	[8]
German army structure (473)	`http://ldf.fi/ww1lod/iga/`
Classes:	ww1:Army, ww1:ArmyCorps, ww1:Brigade, ww1:Regiment, ww1:Division, ww1:Batallion (subclasses of crm:E40_Legal_Body)
Properties:	skos:prefLabel, crm:P107i_is_current_or_former_member_of
Source:	[24], [5]
Other actors (181)	`http://ldf.fi/ww1lod/main/`
Classes:	crm:E21_Person, crm:E40_Legal_Body, crm:E74_Group, org:Post
Properties:	skos:prefLabel, skos:altLabel, org:holds, org:headOf, crm:P107i_is_current_or_former_member_of, schema:colleague, schema:birthDate, schema:deathDate, schema:gender, skos:related
Source:	CU Libraries
Geography of Belgium and France (1312)	`http://ldf.fi/ww1lod/main/`
Classes:	ww1:Country, ww1:County, ww1:Municipality, ww1:Village, ww1:Bay, ww1:Canal, ww1:Cape, ww1:Cove, ww1:Forest, ww1:Hill, ww1:Lake, ww1:Ocean, ww1:Ridge, ww1:River, ww1:Salient, ww1:Valley (subclasses of crm:E53_Place)
Properties:	skos:prefLabel, skos:altLabel, geo:lat, geo:long, crm:P89_falls_within
Source:	IWM Western Front geographical keywords, GeoNames, CU Libraries
Belgian statistical data 1914–1918 (12)	`http://ldf.fi/ww1lod/bsta/`
Classes:	qb:Observation
Properties:	qb:dataSet, geonames:population, sdmxdim:sex, sdmxdim:refPeriod, sdmxdim:refArea
Source:	[1]
Polygons of Belgian provinces 1914–1918 (56)	`http://ldf.fi/ww1lod/bpol/`
Properties:	georss:polygon, geo:long, geo:lat
Source:	HISSTAT (Universities of Ghent, Brussels, and Louvain-la-Neuve, and State Archives of Belgium)

events extracted from Wikipedia's timeline of World War I[11], which are meant to provide an alternate viewpoint into major wartime events. The second is a list of 3604 events that were automatically extracted from a scanned copy of Great Britain. Committee of Imperial Defence [8]. Although these events were harvested from an authoritative source, the entries contain numerous errors due to problems in optical character recognition and structure detection. Domain specialists at CU are currently cleaning up this data and enriching it with with actor and place information; however, this process is ongoing.

[11] `http://en.wikipedia.org/wiki/Timeline_of_World_War_I`

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* [24]. 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 have 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 to the organizations to which they belong.

In addition, the actors were also linked to more general content keywords. However, unfortunately here a misunderstanding regarding modeling caused the actors to be related to the themes (such as "slavery") by the same property (skos:related) not only when they were directly associated with a theme through events, but also when they wrote about those events. Thus, this information should be used with caution.

3.3 Historical Places

Regarding the locations of wartime events, a key point is that events such as battles often refer specifically to geographical features such as ridges or rivers. However, modern gazetteers such as GeoNames mainly contain administrative units (though they may also share a name with nearby geographical features).

Further, whereas geographical features and e.g village level administrative divisions are relatively static for the purposes of the project, larger administrative areas may have changed substantially over the past century. Thus modern gazetteers are not always directly applicable to a historical case like WW1.

To help address these issues, an integrated gazetteer of historical place names was included in the WW1LOD dataset. 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 that was likewise integrated into the dataset. Additional geographical instances also were added manually during the event and actor enrichment process; for example, Pommeroeul (ww1:Village) *crm:P89_falls_within* Hainaut (ww1:Municipality) *crm:P89_falls_within* Belgium (ww1:Country, a subclass of crm:E53_Place, and separate from Belgium the actor of type crm:E40_Legal_Body that holds this administrative area). Coordinate information for these places was sourced through multiple means. First, 1248 contemporary place names were automatically mapped to 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, in the end accepting 836 matches. They also manually located an additional 57 locations.

Finally, 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. To encode these border polygons, the GeoRSS vocabulary[12] was used.

Taken together, the geographical data gleaned from these various sources allows items to be placed in their geographical context using locations and boundaries accurately defined for the war years.

With regard to the rich modeling CIDOC-CRM applies to spatiotemporal entities, it should be noted that again here we have opted for a more simple model. This was possible due to the limited temporal extent of the domain described. However, the vocabulary can easily be extended with the relevant spatiotemporal properties and concepts, should these capabilities be required later.

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 this reason, CIDOC-CRM's temporal representation [4] supports a level of uncertainty in encoding. For instance, it is possible to create a time-stamp 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 flexible timeframes, analyses and visualizations of the temporal relationships between war events can include those with uncertain or ambiguous dates.

Another element that can add to the uncertainty of the temporal extent of wartime events is tendency of actors on opposing sides to regard it differently [8, explanatory notes]. In the case of a battle, for example, one side might include its preparatory attacks to the battle or the consolidation of gains following a successful attack in the timeframe. At present, such differences

[12] `http://www.georss.org/rdf_rss1.html`

are handled in the WW1LOD project merely by expanding the event interval. Formal encoding of such different viewpoints has been been discussed, but remains future work.

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 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 between documents and events 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 [1]. The intended use for these statistics is to enable the identification of correlations between population fluctuations and wartime events. To model these statistics, the project utilizes the W3C Data Cube vocabulary[13].

3.7 Interlinking

While the project strives mainly to include 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 WW1. 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. .

[13] http://www.w3.org/TR/vocab-data-cube/

Table 3 Equivalency links in the dataset

Target	Nr.
IWM and other events (internal)	46
DBpedia events and actors	152
Out of the Trenches events	29
GeoNames places (skos:closeMatch)	836

4 Experience of Ontologies Used

In making modeling decisions, the WW1LOD project has striven to balance the following guidelines: 1) make use of established ontologies (and the CIDOC-CRM in particular) as much as possible, and for incoming sub-datasets: 2) model similar data in a uniform manner, 3) model data in as intuitively as possible, and 4) never lose data included in the original version.

For most sub-datasets, adhering to CIDOC-CRM has fulfilled all of these goals. The first notable deviation was, however, that the decision was made not use CIDOC Appellation resources for labeling entities, instead opting to use the simple and widely used SKOS preferred and alternate label properties. The goal in doing so was to maintain the simplicity of the model, which could be done because none of the data sources contained metadata relating to the appellations themselves.

A second notable deviation was in how relationships between organizations and people are modeled. In principle, CIDOC advocates modeling these also as events and has, for example, activity subclasses for joining and leaving a group (with a wide definition of what constitutes a group, including marriages and families). However, based on prior research [20, 14], this construction is not in line with how people intuitively organize such information. Instead, human end-users generally want to see this data as simple attributes and relations, possibly with attendant metadata such as period of validity. Incidentally, the Getty Foundation also adopted this model for their Thesaurus of Geographic Names and Union List of Artist Names. To an extent, the CIDOC standard itself also acknowledges this perspective by including "shortcut" properties such as crm:P107i_is_current_or_former_member_of that can be used to relate an actor directly to a group.

In the WW1LOD dataset, this property is used where applicable, but the divergent modeling practice also required the definition of additional relationships. For these, the project makes use of multiple generally accepted vocabularies for such data: the W3C Organization Ontology[14], the RELATIONSHIP ontology[15], and

[14] http://www.w3.org/TR/vocab-org/
[15] http://vocab.org/relationship/

the schema.org vocabulary[16]. Particularly used is the facility of the Organization Ontology to define Posts, which can then be linked to the associated people and organizations.

Should the project in the future need a pure event-oriented view of the data, inferencing rules can be put in place to translate between the current model and native CIDOC-CRM [14]. Naturally, the project could have also chosen to encode the data primarily as events and run inferencing the other way when querying, editing, or visualizing, but this would have added a great deal of complexity to common tasks, with no clear benefits to offset the cost.

Another important note regarding modeling is the fact that CIDOC-CRM often contains more specific applicable classes than are actually used in the data. For example, almost all events in the dataset would actually be instances of crm:E7_Activity, which includes all events that are carried out intentionally. Some could be instances of even more specific event subclasses, such as crm:E65_Creation or crm:E6_Destruction.

However, as the sub-datasets have for the most part been automatically converted, the included events were kept on the most general level of crm:E5_Event to preclude introducing any incorrect information into the model. In further manual processing, only some of these have been further refined. This situation is repeated for properties, opting often to use, for example, the neutral crm:P11_had_participant instead of the more specific crm:P14_carried_out_by, which implies agency.

5 Dataset Access

The dataset is published via the Linked Data Finland (LDF.fi) data publication service [10], 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[17], 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 [19]. 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.

6 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 headings[18]. This vocabulary, however, is still preliminary, as the URIs for the added headings 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 contain individual units or events but rather general headings 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 headings 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[19]. 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 stag-

[16] `http://schema.org/`

[17] `http://creativecommons.org/licenses/by-sa/4.0/`

[18] Viewable at `http://skos.europeana.eu/api/find-concepts?q=inScheme:*&rows=1000`

[19] Available at `http://data.aim25.ac.uk/about_t3.php`

ing area at http://ldf.fi/ww1/ for further study and possible integration into WW1LOD.

The Muninn project[20] 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[21], 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. Whereas 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[22] 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.

The 1914 – 1918 Online[23] encyclopedia has collaborated with the WW1LOD project by sharing their own timeline and attendant metadata. These datasets allow sources utilizing the WW1LOD dataset to link to relevant articles on 1914 – 1918 Online.

Finally, Historypin[24] 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.

7 Examples of Use

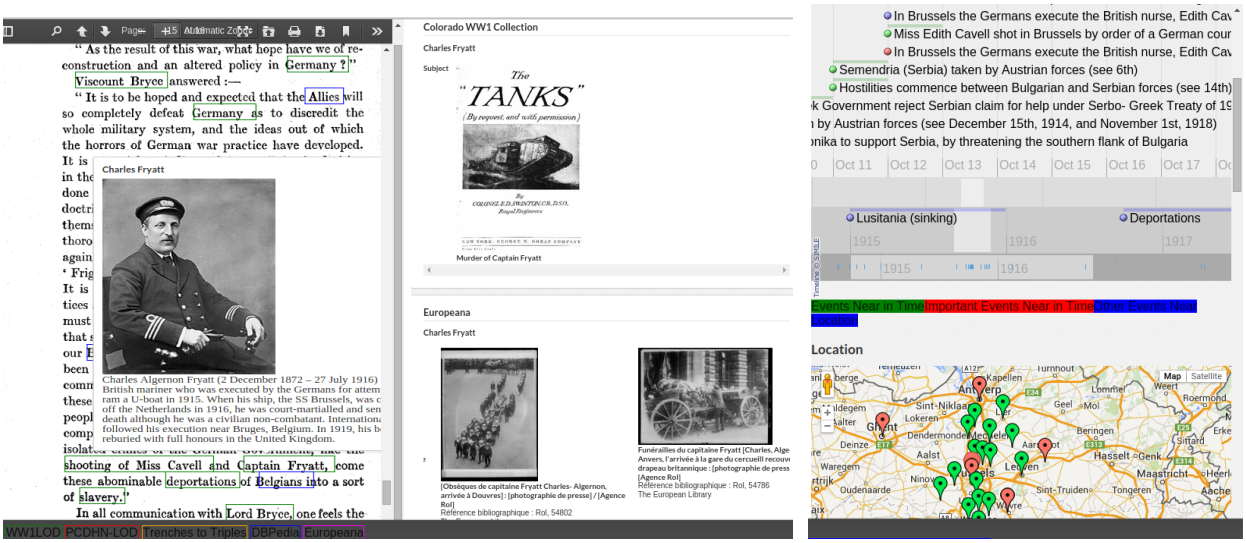
In demonstrating the usefulness of the WW1LOD dataset, a first problem is that only the CU WWI Collection Online has been directly linked to it. The equivalency links to other vocabularies offer little help, due to either a low level of overlap or a high level of generality in these vocabularies. As an example, a SPARQL query[25] for items related to events in West Flanders can be crafted to 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 is demonstrated through a contextual reader prototype[26] [21] developed in tandem with the project. This demonstrator, visualized in Figure 1, uses various methods to overcome issues arising from the lack of detailed annotations in related collections.

First, the application uses dynamic entity extraction to provide context for sources lacking formal metadata. Here, textual mentions of e.g. events, places and people are identified and linked to their context as defined in the associated reference datasets. For the World War 1 configuration, these datasets are WW1LOD, Out of the Trenches, DBpedia, the 1914–1918 Online vocabulary and the Europeana 1914–1918 thesaurus. Being dynamic, this linking can be applied to any PDF or HTML source, be they primary source documents from the Colorado WWI Collection, item pages from Europeana or articles from the 1914–1918 Online encyclopedia. In the interface, the entities extracted are 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 1a with the example of Captain Charles Algernon Fryatt.

Naturally, the automatic extraction of entities can also result in spurious matches and context, for example resulting in the short gloss on "Belgians" being illustrated by a photograph of Jean-Claude Van Damme, as he happens to appear as an example Belgian in the Wikipedia article on them. The extent to which the individual vocabularies used are able to locate entities, as well as the relevance of those entities varies by source and configuration. Generally, in preliminary tests, the Europeana, Out of the Trenches and 1914–1918 vocabularies perform significantly worse than WW1LOD and DBpedia. DBpedia works well for the most notable people and places, but also brings in many spurious

[20] http://rdf.muninn-project.org/
[21] http://www.canadiana.ca/en/pcdhn-lod
[22] http://www.cendari.eu/about-cendari
[23] http://www.1914-1918-online.net/
[24] http://historypin.com/
[25] http://j.mp/1EP1P5i
[26] Available at http://demo.seco.tkk.fi/ww1/



(a) (b)

Fig. 1 The contextual reader interface

matches due to its more general scope. The WW1LOD vocabulary performs well on both precision and recall overall, suggesting the worth of such a specialized vocabulary in interlinking content pertaining to a particular domain. The reasons for this outcome seem to be twofold. First, the vocabulary is more focused and rich in its field. Second, WW1LOD has made a conscious attempt to include alternative names and spellings for concept labels, further increasing recall.

In relation to sources originating from the Colorado WWI Collection, the reader also allows one to evaluate how well the dynamic extraction performs against manual item level indexing of the documents by the CU librarians. Here, preliminary experience seems to indicate that both have their place, but serve different needs. The item level indexing is useful to accurately describe the general content of the document, as well as to link it as a whole to its context. On the other hand, these keywords cannot give context to individual mentions in the source text itself, and neither do they provide the breadth of linkage that the dynamic extraction enables (albeit at the cost of precision). For both uses however, it is apparent that the more focused vocabulary of WW1LOD outperforms the higher level LCSH subject heading classification traditionally used by libraries.

Back in the interface of the reader, clicking on the entity opens a side pane in the interface containing added context. For example, in Figure 1a, the shooting of nurse Edith Cavell is highlighted right next to that of Captain Fryatt's. Clicking on this mention opens the view shown in Figure 1b, where a timeline locates her execution 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. These visual contextualizations make heavy use of the event timelines unified in WW1LOD, and would not be possible without the information contained therein.

This side pane also contains linked content found from multiple sources. These are shown on the right hand side of Figure 1a for Charles Fryatt. Linked content can be brought in from both sources with rich structured metadata, as well as those lacking such. In the example, metadata brings in a relevant primary source from the CU WWI Collection Online. Images of the burial of Captain Fryatt from Europeana, on the other hand, are found not through formal metadata, but rather from a keyword match on his name that appears in the textual description of the images.

The way this works is by the reader extracting from the configured vocabularies a number of relevant textual search terms for an entity of interest and then using the associated collections' existing keyword search functionality to discover related content. Here again the rich data and alternative names and spellings contained in WW1LOD are put to good use. Among the extracted data used in the query are multi-lingual primary and alternative 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 the WW1 configuration, sources queried in this way include WW1 Discovery, Europeana, the Digital Public

Library of America (DPLA), and the European Digital Library.

In such inexact linking, there is again a potential for spurious outlinks. The amount of such depends on both the linking vocabularies and configurations used, as well as the external data source. For example, any items found on WW1 Discovery should be pertinent, but for the other sources this may not be the case. Here, the issue is a matter of calibrating a desired balance of precision and recall, for example by only using the more focused vocabularies for linking, or alternatively by adding external constraints, such as requiring objects matched to originate from the time-span of the war (but thereby missing e.g. later monuments, tributes or writings).

It should also be noted that whether bringing in content through formal metadata or text queries, none of the sources queried make use of CIDOC-CRM as their content organization standard, instead relying on simpler, either Dublin Core or MARC -derived schemata. This again goes to show how in linking together collections, model uniformity is much less important than having a common and detailed enough domain vocabulary.

8 Discussion and Future Work

The WW1LOD dataset provides a rich framework for linking together content 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. The usefulness of the dataset in linking has been demonstrated in a concrete application that is able to bring in and bridge content from a number of sources, both through formal metadata as well as hot linking through using the vocabulary in entity extraction and query expansion.

Relating to CIDOC-CRM, the project highlights how it is the particular instances in a domain that are important to interlinking, while model differences and deficiencies can actually quite easily be traversed. However, in modeling these instances, the standard does seem to provide a good base, both in terms of entity types as well as properties by which to create a rich network of links inside the reference vocabulary. On the other hand, some instances were also identified where the event-based modeling of CIDOC-CRM runs counter to intuition as well as other established standards.

Content-wise, the WW1LOD dataset is currently able to provide a quality-controlled timeline of general events related to the entire war, as well as a much richer representation of a particular subtopic, the occupation of Belgium. As such, WW1LOD currently represents the most comprehensive linked open reference available related to WW1. However, while this is enough to demonstrate the potential of the approach, further work is needed to attain a similar level of quality and detail in additional sub-topical areas. The hope is that researchers who would benefit from inclusion, for example, of geographies for other countries, or the structures of armies other than the German one will add them, thereby enriching the dataset.

Discussions are ongoing with various projects, notably Out of the Trenches, Europeana 1914–1918, DPLA, and 1914 – 1918 Online about 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 [18] to create structured metadata for entities mentioned in the source texts. A similar approach is likely feasible for other mainly textual collections. Future steps the project could take to facilitate this outcome is to reach out to more institutions holding WW1 collections.

Finally, it would be beneficial to formally evaluate how well the vocabulary is able to interlink collections as well as provide context in comparison to other choices. Indeed, plans are under way to do this. Unfortunately, this is a complex undertaking which will take considerable time to plan and execute. First, it is not easy to design a concise test that would measure the degree of support in a meaningful real world scenario. Second, the amount of variables to be controlled for is also high, as the support provided is dependent on details of all the collections and vocabularies linked, as well as a process of tuning the precision and recall of each through configuration.

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. Belgium Ministère de l'Intérieur et de l'Hygiène (1922) Annuaire statistique de la Belgique et du Congo Belge, vol 46. Brussels
2. Binding C, May K, Tudhope D (2008) Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction Via the CIDOC CRM. In: Christensen-Dalsgaard B, Castelli D, Ammitzbøll Jurik B, Lippincott J (eds) Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science, vol 5173, Springer Berlin Heidelberg, p 280–290, DOI 10.1007/978-3-540-87599-4_30, URL http://dx.doi.org/10.1007/978-3-540-87599-4_30
3. Bountouri L, Gergatsoulis M (2011) The Semantic Mapping of Archival Metadata to the CIDOC CRM Ontology. Journal of Archival Organization 9(3-4):174–207, DOI 10.1080/15332748.2011.650124, URL http://dx.doi.org/10.1080/15332748.2011.650124, http://dx.doi.org/10.1080/15332748.2011.650124
4. CRM SIG (2011) How to implement crm time in rdf. Tech. rep., Amsterdam
5. Cron H (1937) Geschichte des Deutschen Heeres im Weltkriege 1914-1918. Biblio-Verlag, Berlin
6. Doerr M (2003) The CIDOC CRM – an ontological approach to semantic interoperability of metadata. AI Magazine 24(3):75–92
7. Giles J (2005) Internet encyclopaedias go head to head. Nature 438(7070):900–901, URL http://dx.doi.org/10.1038/438900a
8. Great Britain Committee of Imperial Defence (1922) Principal events, 1914-1918. History of the Great War based on official documents, London
9. Horne J, Kramer A (2001) German atrocities 1914: a history of denial. Yale University Press, New Haven
10. Hyvönen E, Tuominen J, Alonen M, Mäkelä E (2014) Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. In: [22], pp 226–230, DOI 10.1007/978-3-319-11955-7_24, URL http://dx.doi.org/10.1007/978-3-319-11955-7_24
11. Hyvönen E, Lindquist T, Törnroos J, Mäkelä E (2012) History on the semantic web as linked data – an event gazetteer and timeline for World War I. In: Proceedings of CIDOC 2012 – Enriching Cultural Heritage, Helsinki, Finland, CIDOC, http://www.cidoc2012.fi/en/cidoc2012/programme
12. Lefevre P (ed) (1987-2001) Belgique et la Première Guerre mondiale, bibliographie. Musée royal de l'Armée, Brussels
13. Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes PN, Hellmann S, Morsey M, van Kleef P, Auer S, Bizer C (2015) DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web Journal 6(2):167–195
14. Lin CH, Hong JS, Doerr M (2008) Issues in an inference platform for generating deductive knowledge: a case study in cultural heritage digital libraries using the CIDOC CRM. International Journal on Digital Libraries 8(2):115–132, DOI 10.1007/s00799-008-0034-0, URL http://dx.doi.org/10.1007/s00799-008-0034-0
15. Lindquist T, Long H (2011) How can educational technology facilitate student engagement with online primary sources?: user needs assessment. Library Hi Tech 29(2):224–241
16. Lindquist T, Dulock M, Törnroos J, Hyvönen E, Mäkelä E (2013) Using linked open data to enhance subject access in online primary sources. Cataloging & Classification Quarterly 51(8):913–928, DOI 10.1080/01639374.2013.823583
17. Luyt B, Tan D (2010) Improving Wikipedia's credibility: References and citations in a sample of history articles. Journal of the American Society for Information Science and Technology 61(4):715–722, DOI 10.1002/asi.21304, URL http://dx.doi.org/10.1002/asi.21304
18. Mäkelä E (2014) Combining a REST lexical analysis web service with SPARQL for mashup semantic annotation from text. In: [22], pp 424–428, DOI 10.1007/978-3-319-11955-7_60, URL http://dx.doi.org/10.1007/978-3-319-11955-7_60
19. Mäkelä E, Hyvönen E (2014) SPARQL SAHA, a configurable linked data editor and browser as a service. In: [22], pp 434–438, DOI 10.1007/978-3-319-11955-7_62, URL http://dx.doi.org/10.1007/978-3-319-11955-7_62
20. Mäkelä E, Hyvönen E, Ruotsalo T (2012) How to deal with massively heterogeneous cultural heritage data - lessons learned in CultureSampo. Semantic Web 3(1):85–109
21. Mäkelä E, Lindquist T, Hyvönen E (2016) CORE - a contextual reader based on linked data. In: Proceedings of Digital Humanities 2016, long papers
22. Presutti V, Blomqvist E, Troncy R, Sack H, Papadakis I, Tordai A (eds) (2014) The Semantic Web: ESWC 2014 Satellite Events - ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers, Lecture Notes in Computer Science, vol 8798, Springer, DOI 10.1007/978-3-319-11955-7, URL http://dx.doi.org/10.1007/978-3-319-11955-7

23. Rector LH (2008) Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. Reference Services Review 36(1):7–22
24. Tessin G (1974) Deutsche Verbände und Truppen. Biblio-Verlag, Osnabrück
25. Volz J, Bizer C, Gaedke M, Kobilarov G (2009) Discovering and maintaining links on the web of data. In: Bernstein A, Karger DR, Heath T, Feigenbaum L, Maynard D, Motta E, Thirunarayan K (eds) International Semantic Web Conference, Springer Verlag, Lecture Notes in Computer Science, vol 5823, pp 650–665
26. Waters NL (2007) Why you can't cite Wikipedia in my class. Commun ACM 50(9):15–17, DOI 10.1145/1284621.1284635, URL http://doi.acm.org/10.1145/1284621.1284635