http://conference.ifla.org/ifla78 Date submitted: 30 May 2012

WORLD LIBRARY
AND INFORMATION
CONGRESS:
78TH IFLA GENERAL
CONFERENCE
AND ASSEMBLY

IFLA
2012 Helsinki

LIBRARIES NOW!
INSPIRING
SURPRISING
EMPOWERING

# Leveraging linked data to enhance subject access in online primary sources – a case study of the University of Colorado Boulder's World War I Collection online

**Thea Lindquist**
University of Colorado Boulder
Boulder, CO, USA
E-mail: thea.lindquist@colorado.edu

**Eero Hyvönen**

**Juha Törnroos**

**Eetu Mäkelä**
Aalto University and University of Helsinki
Aalto, Finland
E-mail: first.last@aalto.fi

**Session:** *117* — **Subject access now: inspiring, surprising, empowering** — *Classification and Indexing*

## Abstract:

*Academic users often find work with online primary sources both rewarding and challenging. Improving subject access in these sources is essential as digital collections propagate and work with primary sources becomes increasingly important in humanities curricula. A user needs assessment was conducted with humanities users at the University of Colorado Boulder to facilitate engagement with these sources. Two of the major user needs identified were improving findability and context, particularly for historical subjects.*

*Linked Data can help meet these needs by linking related concepts in the sources using a specialized vocabulary, enriching them with outside resources, and enabling semantically rich services that empower users. This paper discusses a project the authors undertook to enhance subject access in CU's WWI Collection Online by deep linking historical data on the civilian experience in occupied Belgium. This work is intended to lead to a richer understanding of forces shaping the WWI period.*

## I. Introduction

With the wealth of digital cultural heritage collections online, users have greater opportunities than ever before to engage directly with primary sources.[1] Academic users, however, find work with these sources to be both rewarding and challenging. While these digital collections offer extremely valuable resources, particularly for humanities disciplines, studies indicate that they remain underutilized.[2] Improving access for this target group of academic users is therefore increasingly essential. In order to understand their needs and facilitate engagement with the sources, a user needs assessment was conducted with humanities faculty and students at the University of Colorado Boulder (CU). Two of the major needs identified were improving findability and context, particularly for historical subjects.

This paper investigates how Linked Data[3] might meet these user needs using CU's World War I (WWI) Collection Online as a test bed. In addition to representing the collection metadata as Linked Data, our aim is to deep link historical data in the sources related to the civilian experience in occupied Belgium to show the kinds of complex questions that can be answered and automated methods employed in a specialized domain. This approach can help meet user needs by linking related concepts in the sources using a specialized vocabulary, enriching them with outside resources, and enabling semantically rich services that empower users.

## II. Challenges of Working with Online Primary Sources

Work with primary sources has become an increasingly vital component in humanities, and especially history, curricula at the undergraduate and secondary educational levels.[4] Indeed, the use of online primary sources in the classroom is considered fundamental to current pedagogical approaches that encourage critical thinking and inquiry-based, constructivist learning. Lee and Clarke, for instance, explain that "the nonlinear shape of the Web can serve as a lever to encourage students to deal with the multiple sequences, voices, outcomes, and implications of historical narrative."[5] Online primary sources offer distinct advantages over non-digital formats for research as well, mainly in that they are more accessible, searchable, flexible and easily manipulated.

Although the collections digitized and made available by cultural heritage institutions constitute an extremely rich and valuable pool of materials for teaching, learning and research in the humanities, they remain underutilized by both faculty and students for a variety of reasons. Since most of these users rely on Google, the decontextualization of sources, impenetrability of institutional databases, and sheer magnitude of the results all represent major barriers to use. Given the issues with the findability and discoverability of digital primary sources, it is hardly surprising that several studies have identified an increasing demand for easier and more granular subject searching within and among the documents and collections in these databases.[6]

## III. USER NEEDS ASSESSMENT

As a part of an effort to develop a user-centered digital educational tool for work with online primary sources, a user needs assessment was conducted at CU in January 2011.[7] The aim of this research was to understand the needs of academic users in the humanities and facilitate engagement with these sources based on their direct feedback. The study was based on over 20 semi-structured interviews with faculty, graduate students, and undergraduate students representing a range of educational levels and disciplinary areas as well as various degrees of familiarity with primary sources.[8] While academic stakeholders were the focus of the study, the needs they expressed apply, to one extent or another, to all users of online primary sources.

The study confirmed that humanities faculty and students still face significant challenges finding and contextualizing online primary sources.[9] They tend to be unaware of the full range of resources available and deem it inefficient to search multiple databases and websites to identify relevant sources. Once they locate a collection to search, they encounter problems finding and contextualizing individual sources and the information within them, particularly for historical subjects.

Participants reported that bibliographic metadata is often inadequate to expose individual sources and especially sections within them by subject, time period and geographical area with the desired granularity. Since similar concepts are expressed variantly across texts, keyword searching is haphazard. Online primary sources are even more susceptible to decontextualization, since keyword searching encourages users to look for snippets of a document in which a given term is mentioned and then skip forward to the next, rather than reading the document in its entirety.[10] Also, search engines and collections of links to online sources can contribute to this problem by disaggregating individual documents from their archive of origin.

Contextualizing primary sources is often necessary to make them sufficiently accessible for users, especially students and non-experts, to engage with the substance of the material. This context can include displaying the relationships between individual documents as well as resources that help explain how each document, and the information within it, fits into its historical context.[11] Even with relevant sources and adequate context, users may struggle with further challenges inherent to primary-source research: foreign languages, document bias, historical usage, orthography, grammar, paleography/typography, etc.[12] Although all of these issues make it difficult and time-consuming to find and use online primary sources, participants agreed that these sources present a unique educational and research opportunity.

## IV. Advantages of Linked Data

In assessing different options that might meet these user needs, one of most promising was leveraging Linked Data and semantically rich services to increase the interoperability and usability of digital historical collections. According to Tom Heath and Christian Bizer, Linked Data "refers to a set of best practices for publishing and interlinking structured data on the Web."[13] By linking related concepts within and among documents in a way that is understandable to computers, Linked Data allows for (1) the aggregation of data available in distributed online primary sources, (2) new connections to be made among them and visualized in ways that were not previously possible, and (3) the enrichment of the data through links to outside resources like DBpedia.[14]

Digital cultural heritage content, and historical material in particular, presents a level of complexity that can benefit from semantically enriched (meta)data and intelligent user services, both in improved findability and enriched context.[15] They expose the complex, often nonlinear relationships among the topics, people and places buried within the sources, particularly when drawing on ontologies and other specialized vocabularies that impart meaning to these concepts and the relationships among them in a given historical domain. Linked Data applications can also help mitigate common issues with formulating searches based on subject headings and with keyword searching in historical sources, for instance by suggesting search terms to users that help them to refine and focus their searches. In these ways, they facilitate effective subject access to historical content.

Linked Data, moreover, allows for the richer contextualization of sources by making connections not only within collections but also to relevant outside sources, thus also enabling interoperability, sharing and reuse of data across historical collections. Additionally, it can be presented in a way that exposes the organizational structure of collections, which not only preserves the original context of the documents but also alerts users to the types of materials available. Taken together, these improvements can help overcome the well-known limitations of subject headings and imprecision of keyword searching, facilitate comparisons of historical data across space and time, and enable work across multiple collections. Further, Linked Data is easily consumed and encourages the development of intelligent applications that are easy-to-use and present the user with a range of options for analyzing and visualizing the data.

## V. WWI Linked Data Project

This project involves computer scientists from Aalto University[16] and a library subject specialist/domain expert from CU. The primary dataset used is CU's WWI Collection Online, which comprises over 1,100 titles (55,000 pages) published from 1829 to 1922, with the vast bulk of the material published between 1914 and 1918.[17] The provenance of the collection is not entirely clear, but it likely entered the holdings of the CU University Libraries in the 1920s or 1930s by way of the Colorado in WWI Project, which History professor James Field Willard undertook to document citizen and state activities during the war.[18] The collection publications originate mainly from the U.S. and touch on a variety of geopolitical regions and topics, from ethnic and religious conflict to empire and colonies. A range of genres is represented including pamphlets, books, reports, speeches and maps. Negotiations are currently underway to publish this content as Linked Open Data under a Creative Commons 2.0 license.

One of the project's main objectives is to enhance subject access in the online collection and create context for the documents by establishing links between data points in the collection, datasets incorporated into the project, and external data sources like DBpedia and Freebase.[19] Another is to facilitate the annotation of and deep linking of concepts among WWI collections in a specialized historical subdomain, in this case pertaining to the civilian experience in occupied Belgium in WWI. This topic was selected not only because it was well-represented in the collection but also because the impact of "total war" on civilian populations is a current area of scholarly interest. Most of the publications falling into this category deal with the hardships Belgians suffered during the German invasion and occupation, particularly atrocity incidents such as killings and worker deportations and the impact of military rule on day-to-day life. Among the datasets converted to RDF (Resource Description Framework - Linked Data format) thus far are the collection metadata (MARC), standard vocabularies from the Imperial War Museum (IWM),[20] information on German atrocities in Belgium, and the German army hierarchy.[21]

Deep linking using a specialized vocabulary on WWI Belgium combined with an intelligent user interface is designed to demonstrate the types of complex questions that can be answered to meet user needs in this subdomain, such as: Is the scale of the atrocity incidents involving German troops in Belgium accurately reflected in the collection literature? What divisions of the German army were involved in the most incidents? What was the geographic distribution of deportations from the Belgian provinces? This type of functionality is intended to lead to a richer understanding of the many forces shaping the WWI period. Given the very specific nature of this subdomain and the lack of existing ontologies, this vocabulary needed to be created, adapting terminology and structures from the standard print bibliography on Belgium in WWI,[22] incorporating feedback from historians in the field,[23] and linking wherever possible to relevant terms from other datasets included in the project, e.g., the collection metadata and IWM vocabularies.

Our work utilizes the existing FinnONTO ontology system[24] and extends it with annotations on the datasets mentioned above and a general event-based framework for WWI currently in process using the SAHA semantic annotation tool.[25] Additionally, we are employing the ARPA tool[26] to automate part of the annotation process. ARPA is an information extraction tool that automatically mines named entities and keyword concepts from ontologies in textual documents. The suggested annotations can then be validated and corrected manually using the SAHA editor. Finally, development of a customized WWI web portal based on the faceted HAKO portal engine[27] is underway to facilitate searching and browsing the data by topics, people, places and time periods, and to represent it in visual and interactive ways.

The WWI framework and other structures we have created are meant to be shared, thus providing the “semantic glue” that binds separate datasets together and allows searching and browsing between them. Moreover, the strategy we have developed for this project is intended to be adaptable to other historical domains and datasets, particularly those related to conflicts like the U.S. Civil War or World War II.[28]

## VI. Conclusion

By linking related concepts across WWI datasets using a specialized vocabulary and enabling semantically rich services, we hope to empower users to find and use online primary sources efficiently and effectively. The upcoming centenary of the War will undoubtedly generate much interest, especially in the countries that were involved. We can use this moment to engage users actively with the past and with the wealth of digital materials that cultural heritage institutions have made available.

**Acknowledgments** This work has been supported by the Fulbright U.S. Scholar Program, Finnish Cultural Foundation, and Tekes – the Finnish Funding Agency for Technology and Innovation.

[1] Primary sources are documents, objects or other evidence about the past that were created during the time period the historical events took place or by those who experienced those events. Some examples include: diaries, letters, speeches, government documents, books, interviews, photographs, audio and video recordings, and artifacts.

[2] D. Harley, “Use and Users of Digital Resources: A Survey Explored Scholars Attitudes about Educational Technology Environments in the Humanities”, *Educause Quarterly* 30, no. 4 (2007): 12-20.

[3] T. Heath and C. Bizer, *Linked Data: Evolving the Web into a Global Data Space*, Synthesis Lectures on the Semantic Web: Theory and Technology, ed. J. Hendler and F. van Harmelen (San Rafael, CA: Morgan & Claypool, 2011). Freely available at: http://linkeddatabook.com/editions/1.0/.

[4] See, e.g., J.K. Lee, “Digital History and the Emergence of Digital Historical Literacies”, in *Technology in Retrospect: Social Studies in the Information Age, 1984-2009*, ed. R. Diem and M.J. Berson (Charlotte, NC: Information Age Publishing, 2010), 78-80, and D. Malkmus, “’Old Stuff’ for New Teaching Methods: Outreach to History Faculty Teaching with Primary Sources”, *portal: Libraries & the Academy* 10, no. 4 (2010): 414-416.

[5] J.K. Lee and W.G. Clarke, “High School Social Studies Students’ Uses of Online Historical Documents Related to the Cuban Missile Crisis”, *Journal of Interactive Online Learning* 2, no. 1 (2003): 3.

[6] M.C. Pattuelli, “Modeling a Domain Ontology for Cultural Heritage Resources: A User-Centered Approach”, *Journal of the American Society for Information Science & Technology* 62, no. 2 (2011): 314-342.

[7] CU is a Carnegie Research University (very high research activity) with a range of master's and doctoral degree-granting programs in the humanities.

[8] Participants were rostered in seven humanities colleges and departments on campus: Architecture and Planning, Classics, English, French and Italian, History, Music, and Religious Studies, all of which offer doctoral-level programs.

[9] This summary does not include the full range of user needs identified by the study, but rather just those that relate to the topic of this paper. For a full discussion, see T. Lindquist and H. Long, "How Can Educational Technology Facilitate Student Engagement with Online Primary Sources?: A User Needs Assessment", *Library Hi Tech* 29, no. 2 (2011): 224-241.

[10] J. Garrett, "KWIC and Dirty? Human Cognition and the Claims of Full-Text Searching", *Journal of Electronic Publishing* 9, no. 1 (2006), available at: http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;cc=jep;q1=garrett;rgn=main;view=text;idno=3336451.0009.106 (accessed 13 February 2012).

[11] One student gave the following example of the type of context she would like to see: "I would love more background and context for the primary sources I work with. [Many online collections of primary sources] just present the source but give no sense of whether a letter, for instance, was delivered in middle of cholera outbreak." (Lindquist and Long, 233).

[12] T. Lindquist and H. Wicht, "Pleas'd By a Newe Inuention?: Assessing the Impact of Early English Books Online on Teaching and Research at the University of Colorado at Boulder", *The Journal of Academic Librarianship* 33, no. 3 (2007): 347-360.

[13] Heath and Bizer, chap. 2, "Principles of Linked Data", accessed 23 May 2012: http://linkeddatabook.com/editions/1.0/#htoc8.

[14] DBpedia is a Linked Data version of Wikipedia (http://www.dbpedia.org/).

[15] E. Hyvönen, "Semantic Portals for Cultural Heritage", in *Handbook on Ontologies*, 2d ed., ed. S. Staab and R. Studer, International Handbooks on Information Systems (Berlin: Springer, 2009).

[16] Semantic Computing Research Group, see http://www.seco.tkk.fi/.

[17] See http://libcudl.colorado.edu/wwi/index.asp.

[18] These materials provided the basis for the University of Colorado Historical Collection and in turn the University Archives (David M. Hays, "The History of the Archives, University of Colorado at Boulder Libraries, 1917-2011" [unpublished paper, Archives, University of Colorado Boulder Libraries], 1-2).

[19] See http://www.freebase.com/.

[20] These are approved event keywords relating to WWI, approved geographical keywords relating to the Western Front based on the Getty TGN taxonomy and extended by terms related to IWM collections, and the IWM's theme taxonomy with coverage for WWI. Thanks are due to the Imperial War Museum for sharing these vocabularies.

[21] Special thanks to John Horne and Alan Kramer of Trinity College Dublin, who gathered and analyzed the atrocity data and granted permission to include it in the project (J. Horne and A. Kramer, *German Atrocities, 1914: A History of Denial* [New Haven: Yale University Press, 2001], Appendix 1, 435-439).

[22] P. Lefèvre and J. Lorette, eds., *La Belgique et la Première Guerre mondiale: Bibliographie*, 2 vols. (Brussels: Musée Royal de l'Armée, 1987-2001).

[23] Many thanks are due to Martha Hanna (University of Colorado Boulder), Sophie de Schaepdrijver (Pennsylvania State University) and Tammy Proctor (Wittenberg University) for their suggestions.

[24] See http://www.seco.tkk.fi/projects/finnonto/.

[25] For publications and source code download, see http://www.seco.tkk.fi/services/saha/.

[26] See http://www.seco.tkk.fi/services/arpa/.

[27] See http://www.seco.tkk.fi/tools/hako/.

[28] The nature of the content facilitates connections based on military concepts and structures like branches of service, regiments and battles.