Bridging Communities of Practice: Emerging Technologies for Content-Centered Linking

Douglas W. Oard, Amalia S. Levi, Ricardo L. Punzalan
College of Information Studies, University of Maryland, College Park, MD, USA

Robert Warren
Big Data Institute, Dalhousie University, Halifax, NS, Canada

Abstract

This paper describes the potential of new technologies for linking content among cultural heritage collections and between those collections and collections created for other purposes. In recent years, museum professionals, archivists, librarians, and digital humanists have been working to render cultural heritage metadata in an interoperable form as linked open data. Concurrently, computer and information scientists have been developing automated techniques that have significant implications for this effort. Some of these automated techniques focus on linking related materials in more nuanced ways than have heretofore been practical. Other techniques seek to automatically represent some aspects of the content of those materials in a form that is directly compatible with linked open data. Bringing these complementary communities together offers new opportunities for leveraging the large, diverse and distributed collections of computationally accessible content to which many of us are now contributing.

Introduction

If the three V's of "*volume*, *velocity*, and *variety*" is the mantra for "big data" today, we might coin three D's of "*dispersed*, *described*, and (increasingly) *digitized*" as characterizing some of the challenges and opportunities for the cultural heritage collections of today. Digitization affords us with opportunities for unprecedented flexibility for access and use. Use now involves more than just reading or viewing, however; it can also involve giving back. In particular, our users can describe what they find in ways that they hope will have meaning to others. Digitization is one part of this revolution in description, dispersion is the other. Importantly, the users of our collections can describe things that are not now, and perhaps never were, together. In other words, they, and we, can create not only descriptions of things we have, but also descriptions of relationships between things, regardless of where they are. Adding to both the promise and the complexity of this new world, one person can describe the descriptions of another.

None of this is new; it is a world that we have been living in for some time. What is new is that powerful new computational tools have in recent years been developed that can help us to harness this new potential. These opportunities arise from two long-standing investments in computer science and allied disciplines: techniques for linking content, and techniques for building linkable content representations. The first of these is known by the rather unlikely name *wikification*, reflecting its roots as an abstract technical task; the second is the basis for what we now call Linked Open Data (LOD). We have already seen some crossover between cultural heritage professionals and these technologists, but there is much more to be done. In particular, there has been considerable buzz around the use of LOD in Libraries, Archives, and Museums (LAM) in recent years, which has led to the emergence of a vibrant LODLAM community. In this paper, we seek to take the next step, connecting that community and others with what we

somewhat tongue in cheek have referred to elsewhere as New Useful Technical Services (NUTS). In an effort to adopt a somewhat more academic tone, we will in this paper refer to those new capabilities simply as the contributions of computer science.

This paper is organized as follows. First, we provide a brief overview of current work in LODLAM. This is followed by a review of recent research in computer science on the development of automated (or machine-assisted) technologies for content linking and the creation of linkable content representations. We then proceed to identify and describe some of the opportunities, challenges, and strategies for leveraging these capabilities in the heritage field. Finally, we look to the future, offering suggestions for how museums, archives, and libraries can harness as well as influence the present and future directions of this work.

Linking Cultural Heritage

Heritage professionals, administrators and scholars are making considerable investments in linked data as a strategy to transcend the "silos" of content. To date, a principal focus of these efforts has been on publishing structured data in increasingly accessible ways by improving and sharing bibliographic and archival data and on promoting interoperability through datasets, element sets, and value vocabularies (Keller et al., 2011; NISO, 2012). A significant challenge is the fact that cultural heritage encompasses both tangible and intangible expressions and products of cultures. As a result, heritage collections are dispersed, heterogeneous, and subject to (mis-) interpretation. Moreover, as much of the effort to date has focused on data and metadata found in repositories, linking to cultural products beyond the reach of those repositories has received less attention.

Several initiatives now point to the desire for greater connectivity, accessibility, use, and exchange of cultural heritage data. The Museum API wiki (http://museum-api.pbworks.com), for instance, lists APIs and machine-readable sources in about 70 museums, libraries and archives. Its "Cool stuff made with cultural heritage APIs" enumerates several examples of creative uses of openly available cultural data. In addition, aggregating LAM collections has been recently promoted as one way of partially mitigating the dispersion of cultural heritage holdings; ArchiveGrid (http://beta.worldcat.org/archivegrid/), the Open Archives Initiative (OAI) (http://www.openarchives.org/), and Social Network of Archival Context (SNAC) (http://socialarchive.iath.virginia.edu/) are prominent examples. OCLC's ArchiveGrid aggregates archival material from thousands of institutions worldwide. OAI encourages open data among archives through standards for metadata interoperability. SNAC, a collaboration of the University of Virginia, UC Berkeley and the California Digital Library, uses the EAC-CPF standard to aggregate distributed historical records. To these examples we might add others, including the Amsterdam Museum, which makes use of linked open data to make its collection available on the Web (http://datahub.io/dataset/amsterdam-museum-as-edm-lod).

LOD initiatives in the humanities have to date been shaped by an emphasis on leveraging existing metadata, a natural choice given the substantial quantities of high-value metadata and the useful structure present in much of it. In 2011, the NEH-funded LODLAM summit (http://lodlam.net/) brought together LOD innovators from libraries, archives and museums, and since then it has been active in disseminating LOD resources and research. Other groups, notably including the W3C Linked Data Incubator Group (http://www.w3.org/2005/Incubator/lld/), ALA's Library Linked Data Interest Group (http://www.ala.org/lita/about/igs/linked/lit-iglld), the Library of Congress' Bibliographic Framework for the Digital Age (http://www.loc.gov/bibframe/), and the NEH-funded Linked Ancient World Data Institute

(http://wiki.digitalclassicist.org/Linked_Ancient_World_Data_Institute) also focus principally on metadata.

Achieving interoperability requires significant coordination effort. Prominent examples of large-scale humanities projects employing LOD include PELAGIOS (http://pelagios-project.blogspot.com/), which aims to help scholars use ancient world data in meaningful ways, Civil War Data 150 (http://www.civilwardata150.net/), which utilizes structured data across state libraries, archives and museums to promote sharing Civil War data, Muninn (http://datahub.io/dataset/muninn-world-war-i), which utilizes both structured and unstructured data from different archives to promote sharing Great War data, Linked Jazz (http://linkedjazz.org/), which aims uncover meaningful connections between documents and data related to the personal and professional lives of musicians, and Linking Lives (http://archiveshub.ac.uk/linkinglives/), which has developed an end-user interface using linked open data derived from Archives Hub (http://archiveshub.ac.uk/), a gateway to the datasets of 180 institutions across the UK. Europeana (http://pro.europeana.eu/linked-open-data) and the Digital Public Library of America (http://dp.la) are well-known examples of institutional interoperability. CultureSampo (http://www.kulttuurisampo.fi/) is a semantic platform with similar goals for Finnish culture. The recent introduction of Interactive Data Transformation tools, such as Google Refine (http://openrefine.org/), used by projects such as "Free your Metadata" (http://freeyourmetadata.org/; van Hooland et al., 2013), offers some potential to accelerate such projects.

We are interested in much more than "institutionalized content," however. Members of the public routinely generate content, such as blogs, wikis or tweets, that can complement, extend, or provide additional perspectives on the content of LAM collections. Although computationally malleable, such content is less often drawn upon, in part because its associated metadata is not always interoperable with cultural heritage datasets, and perhaps also because we still think of the Web more as a dissemination channel than a resource.

Machines That Learn From Us

Computer science research on what we might call content linking technologies has received hundreds of millions of dollars of investment over the last few decades. Machines can now learn to perform well-structured tasks by observing how people perform those tasks, an approach referred to as machine learning. Over the past half-decade or so, computer scientists have applied machine learning to two tasks that have significant implications for our work with cultural heritage collections: linking content, and linking data that is automatically extracted from (or inferred about) that content.

To this end, computer scientists have been developing two novel technologies that can directly leverage content, not just structured metadata: *wikification* (Mihalcea & Csomai, 2007) and (the somewhat mis-named) *information extraction* (Milne & Witten, 2008). *Wikification* automatically builds hypertext links inside of previously unlinked content, earning the name because the earliest systems in this line of work learned from the way in which people have built links inside Wikipedia pages. *Information extraction* is the more ambitious task, seeking to automatically construct linked data from unstructured content (NIST, 2012). At present, both of these technologies are most capable when applied to machine-readable text, although there has also been some work on speech, audio more generally, images, and video. As is typical early in the technical development life cycle, initial research on these questions has focused on the content that is both easily available and reasonably representative of some real problem, which

of course in not necessarily the content that will ultimately be most important to any specific users of the technology (He, et al., 2011).

The research on *information extraction* emerged from *computational linguistics*, one of many subfields in computer science. In 2007, researchers in another subfield, *information retrieval*, began exploring techniques for the considerably simpler *wikificatio*n task, which extended an earlier line of work on automating construction of hypertext that dates to the dawn of the Web (Meij, et al., 2011; Mihalcea & Csomai, 2007). Developments in these technologies have resulted from a concerted worldwide effort to develop content analysis and linking technologies. The core technology dates to 1987, when the Message Understanding Conference first focused on an *entity detection* task in which the goal was to automatically detect specific references to named entities (people, places, and organizations) and other specific content (e.g., dates) in unstructured text. By 1999, techniques for *entity detection* were sufficiently mature that the more ambitious task of *entity tracking*, in which mentions of the same entity in different documents were to be detected, could be undertaken in the Automated Content Extraction evaluations. A decade later, *entity linking* techniques were sufficiently mature that the even more ambitious *machine reading* task, in which the goal was to automatically create linked data directly from the content of digital documents, could be undertaken in various venues, most recently, in the Text Analysis Conference (McNamee, et al., 2011; NIST, 2012).

In a world in which links between content and between data derived from or describing content are plentiful, we should not have been surprised to see the emergence of the third core technology: tools for reasoning over the resulting so-called *knowledge graphs*. For example, it is possible today for a machine to read in one graph (e.g., the graph of communication patterns in an e-mail collection) and to write out another (e.g., reporting relationships in an organizational hierarchy) (Diehl et al., 2007). While such tools are far from perfect, they are able to work at scales far larger than could any individual scholar, or even any team of scholars. In a world where we can automatically construct structured data from content at an unprecedented pace, the potential of graph manipulation tools that can transform that content in ways that are relevant to the needs of scholars, and the interests of the broader public that our scholarship serves, is significant.

Bringing Communities Together

Building bridges between the vibrant LODLAM community and the relevant computer science communities that can bring potentially transformational capabilities to that undertaking begins with dialogue and exchange of ideas. With that goal in mind, we organized a two-day workshop in September 2013 at the University of Maryland, College Park. The workshop sought not to answer specific questions, but rather to help further the process of asking questions that might ultimately have the potential to transform the way we think. The two dozen participants included cultural heritage scholars and professionals, digital humanities scholars, and computer scientists. The goals of the workshop were to: (1) draw on multiple perspectives to identify important opportunities that no one community or researcher could identify alone, (2) generate ideas that would ultimately be shared with others, and (3) envision next steps to weave these communities more closely together.

Workshop sessions were organized to maximize time spent on interaction among participants rather than on creation of refined products. Introductory presentations outlined current capabilities and limitations of LODLAM and relevant technologies from computer science. Early discussions focused on the brewing paradigm shift described above. A sequence of visioning

exercises and breakout sessions brought together groups to document potential employment scenarios for specific technologies, and to identify possible ways of leveraging the resulting opportunities. In this section, we describe our personal interpretation of the outcomes of that discussion.

Communities: Forging a Common Ground

A symbiotic relationship: The seemingly discrete communities that we have brought together are in fact closely linked. The old notion that LAM institutions simply provide content, that computer scientists simply develop computational techniques, and that digital humanities scholars are the ones who seek, harvest, and interpret content is at best one view of what is in reality a complex and interlinked tapestry. Linked data and linked content are boundary objects, central to each of these communities, and each develops ways of interacting with those boundary objects. There are already a few cross-fertilizations among these fields and our challenge is to create more such opportunities

A two-way value proposition: We must be able to articulate not only why LOD are valuable for humanities, but why they are valuable for computer science as well. There is a gap between the abstracted tasks that computer science research starts with and the real issues that LAM face. Despite the great interest in LODLAM, creation of the robust shareable datasets that computer scientists would need to guide their experimentation remains something of a challenge. LAM have content that presents challenges (e.g., non-digitized content), and they need to enumerate things, digitize them, transcribe, edit, and annotate them, in ways that will make what they have more computationally malleable. Current methods in computer science have their limitations too: for example, they work well with clean content (e.g., a well-written, online NY Times article), but do poorly with more challenging data (e.g., speech, tweets, manuscripts). Computer science needs to understand the implications of these limitations on the cultural heritage sector; if the resulting tools are not robust, that will limit the ways in which they can be used.

The need for small projects: There is a need to try many small things, and then think about how to build reusable infrastructure from those that turn out to make a real difference. The key to success when building things is not necessarily to build the right thing first, but rather to build many things quickly, fail early, and learn as you go. In a world in which resources will always be limited, we must therefore learn to do many things at once with the resources that we have, and this means that we will need to conceptualize and conduct many small projects, the results of which can guide our thinking.

Evaluation: Computer scientists measure the accuracy of the decisions their system make, but ultimately it is the effects of those errors, not their mere existence, that we care about. Moreover, computer scientists typically measure how well their tools do on data similar to what they saw during training, and the real world is of course considerably messier than that. Evaluation resources generally, and evaluation measures in particular, are thus another boundary object between our communities. Because computer scientists are to some extent data agnostic (i.e. they work with the data that is available to them), it is incumbent on the LAM stakeholders and humanities scholars in this collaboration to learn to help guide the work of the computer scientists in productive directions by working with them on selecting cultural heritage materials that can offer useful insights and on designing evaluation measures that reflect the results that are most needed.

Reward structures: Different communities work within different incentive structures. It is important to understand the degree to which these structures are consonant with an eye towards

possibly instigating improvements where needed. Workshop participants discussed the disparity of funding between the sciences, humanities, and LAM. One suggestion is for LAM, digital humanities and computer science communities to team up in solving common cultural heritage data problems. To this end, some suggested creating tool suites that would be enabling (i.e. that would promote capability, and not just the production of artifacts). Others suggested projects with multi-source application problems that would involve images and text and could have multi-phase solutions.

Technology: From Data to Knowledge

The knowledge graph: By encoding aspects of our interpretations as graphs that encode relationships between "things," we gain access to a powerful new mode of expression whose power arises from the machine's ability to explore for patterns and our human ability to interpret the patterns that are found, thus potentially further enriching the knowledge graph. Of course, graphs are but one way of encoding knowledge, and they will surely be better suited to some uses than others. Nonetheless, the iterative combination of machine and human reasoning, layering interpretation over facts, offer us new ways of thinking, acting, and collaborating.

The primary artifact for linking: Different stakeholders will naturally seek to link different things. Some focus on linking objects, some on linking descriptions of objects, and some work at even high levels of abstraction, linking ideas, and linking conversations about those ideas. While we benefit from talking together to see what is common across these settings, we must also bear in mind that we are not all trying to do quite the same thing. There is strength in diversity, and we should not strive for complete convergence, at least not at this point in our thinking about these challenges and opportunities.

Multi-perspective LOD: Computer Science, and Library Science, tend to view data (and metadata) as fixed and objective, but humanists are interested in cultural biases inherent in the data. We should find ways to have more nuanced metadata, including platforms that will support identifying and analyzing such biases, and will allow for cultural constructs to be integrated in LOD. At the heart of LOD is that "authority" is derived from the source of the description rather than that which is being described, allowing for multiple concurrent interpretations of the content to exist, to be found, and to be further interpreted.

A big technology tent: We will also need to involve a far broader range of technologists in the discussion than we have to date, including experts in digital imaging and processing, multimedia information systems, optical character recognition, and human-computer interaction.

User-centered design: Interestingly, both computer science and LAM institutions both see themselves as existing to serve users, but they have different users in mind. Left to their nature, computer scientists will build a "Field of Dreams" based on what they thing their users want. But in the context we are discussing, it is the LAM institutions who are users of what the computer scientists will build. Of course, LAM institutions too exist to serve their "users," and those users are the real people this entire complex ecosystem exists to serve. If we are to build what the real users need, we need LAM institutions to function as translators between the ultimate users and the computer scientists that are seeking to serve real needs. And to the extent that we see humanities scholars as among our ultimate "customers," we will want them at the table as we think through these issues.

Implications for Cultural Heritage Institutions

Item-level description: Content linking technologies such as those described above should be of particular interest for museum collections. Library collections are typically dominated by print materials that are available in many institutions globally, whereas archival collections are dominated by materials described at an aggregate level (i.e., at the fonds, series, or folder level). Museum collections, though, consist of unique items described at item level, which is a sweet spot for creation of linked data.

Digitization vs. description: At this point in our evolution, often much more of our collections have been described than have been digitized. Moreover, the uptake of the "More Product, Less Process" method (Greene & Meissner, 2005) in archives is thinning out some of the description that in earlier times might have been created. To the extent that content-based linking proves useful to LAM institutions, this could incentivize additional digitization as a cost-effective way of generating item-level description. The imperative to serve human viewers by "putting things online" was a powerful early incentive for digitization; in an odd twist, perhaps we will find that satisfying the voracious appetites of our own machines will become one motive force (among many) in the next step of our evolution.

User-contributed context: New kinds of materials, born of the Web, are starting to augment what memory institutions have traditionally collected (e.g., e-mails together with letters, blogs together with diaries). Indeed, to the extent that online content exemplifies the multi-directionality of human experience and memory (Rothberg, 2009) we welcome this recent development. But the online world can contribute more than just content; it can contribute as well to helping to make sense of this content. Allowing users to make their own links among collections through entity linking promotes the performative interpretation of an object's meaning and value based on a user's perspective and worldview. In addition, it allows humanities scholars to "read" objects positioned in a network of other objects at a specific time (Drucker, 2013). Users want to be able to make connections easily and seamlessly. Technologies that support this are recurring themes in Museums and the Web conferences (example, Straup Cope 2009; van Dijk, Kerstens & Kresin, 2009; Miller & Wood, 2010). LOD enables museums to participate in "mash-up" culture while retaining a reference to the original source. Cultural heritage institutions are coming to terms with the fact that once content goes online, they share control with the users of that content (Adair, Filene & Koloski, 2011).Increasing however re-use of and reference to their assets can help museums to benefit society in new ways.

Next Steps

Talk is one thing, action another. We conclude by looking to the future. We see at least three broad themes that deserve our attention.

Institutional Structure

Bridging humanities and computer science research makes it possible to do some things we want to do, but also some things for which we might not yet have thought through all of the consequences. To understand implications of this intervention, we will need to educate a new generation of scholars who are adept in thinking in both ways, and our institutions will need to evolve to create places where that new generation can be nurtured, mentored, and supported, and where they can further pass on what they have learned to a next generation that will ultimately step forward to take their place. It is not yet clear what these institutions will look like, now what these working styles will be, but we can already see some points of intersection emerging in the field that calls itself digital humanities, and few things will be more important to the future

of this endeavor than learning from those experiences and ultimately getting these institutional design issues right.

Collaboration

Looking to the far future is all well and good, but every journey begins in the present and in our present we must learn to collaborate effectively across established disciplines if we are to begin to build the bridges that we need. Here, there are two fundamental issues. At the most basic level, we need awareness of what might be done. Workshops like the one we have described, papers such as this one, and hackathons such as the one we plan to run at code4lib in Raleigh, North Carolina, between this writing and the time this paper is delivered, are useful first steps. Ultimately, we will need to develop ways of institutionalizing the process of developing shared visions. The second key issue is that we must learn to bridge a variety of subtle cultural differences between our research communities. Andreas Paepcke (2008) has written cogently on some aspects of this from the perspective of a computer scientist. We could benefit from sharing additional perspectives on these important issues.

Research Support

To quote a phrase made popular by the movie *The Right Stuff*, "No bucks, no Buck Rogers." Or to draw on management theory, if you want to know what an organization truly values, you should look not to policy statements, but to its budget. Ultimately, progress in any endeavor depends on resource allocation, precisely because resources are always limited. This was, therefore, a natural focus for our final discussions at the workshop. We have four broad themes to recommend. First, look for leverage. The large investments being made in computer science are driven by a diverse set of needs, including health care (e.g., NIH), security (e.g., DARPA), and curiosity (e.g., NSF). Any of these sources might generate further advances that we can leverage, and some (particularly curiosity-driven computer science) might draw inspiration from the research questions of LAM institutions and of humanities scholars. Leveraging investments being made for other purposes will only get us so far, however, so step two is to encourage our funding agencies to come together around joint programs of mutual interest. Here, the Digging into Data, with collaborative funding from NEH, IMLS, and NSF, is a wonderful example. Each funding agency can, and must, act in furtherance of its own mission, but for problems that demand collaborations this broad joint funding offers important potential. Third, build boundary objects. Computer scientists have an insatiable appetite for collections that capture some essence of real problems that are worth solving. Working together to select and assemble such collections can be a worthwhile endeavor. Finally, start small and fail often. This will sound like strange advice when resources are limited. But remember, if we knew what we were doing, we wouldn't call it research.

Acknowledgments

This work has been supported in part by a Digital Humanities Startup Grant HD-51766-13 from the National Endowment for the Humanities. We are grateful to the other participants in the September workshop for generously sharing their insights and perspectives: Brett Bobley, Travis Brown, Perry Collins, Dina Demner-Fushman, Tim Finin, Kevin Ford, Ophir Frieder, Lise Getoor, Corey Harper, Heng Ji, Trevor Muñoz, Christina Pattuelli, Rob Sanderson, Ryan Shaw, Paul McNamee Boyan Onyshkevych, Jon Voss, Gunter Waibel and student volunteers Zahra Ashktorab, Ning Gao, Rashmi Sankepally, Jyothi Vinjumur and Tan Xu. Without them, this paper would not have been possible. We emphasize, however, that our reporting in this paper of what we learned is based on our interpretations, and does not necessarily reflect the views of any other specific workshop participants, and moreover that any views, findings, conclusions, or recommendations expressed in this paper do not necessarily represent those of the National Endowment for the Humanities.

References

Adair, B., Filene, B. & L. Koloski, Eds., (2011). *Letting Go?: Sharing Historical Authority in a User-Generated World.* Philadelphia, PA: Pew Center for Arts & Heritage.

Cope, A.S. (2009). "The Interpretation of Bias (and the Bias of Interpretation)". In J. Trant and D. Bearman (Eds). *Museums and the Web 2009: Proceedings*. Toronto: Archives & Museum Informatics. Published March 31, 2009, consulted February 24, 2014. http://www.archimuse.com/mw2009/papers/cope/cope.html

Diehl, C. P., Namata, G. & L. Getoor, (2007). "Relationship Identification for Social Network Discovery". *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*. Vancouver, BC: AAAI Conference. 546-552.

Drucker, J. (2013). "Performative Materiality and Theoretical Approaches to Interface". *Digital Humanities Quarterly* 7(1). Last updated, January 30, 2014, consulted February 24, 2014. http://www.digitalhumanities.org/dhq/vol/7/1/000143/000143.html

Greene, M., & D. E. Meissner, (2005). "More Product, Less Process: Revamping Traditional Archival Processing". *American Archivist 68(2)*, 208-263.

He, J., de Rijke, M., Sevenster, M., van Ommering, R., & Y. Qian, (2011). "Generating links to background knowledge: a case study using narrative radiology reports". *Proceedings of the 20th ACM international conference on Information and knowledge management CIKM'11*. New York, NY: ACM. 1867–1876. doi:10.1145/2063576.2063845.

Hyvönen, E., et al., (2009). "CultureSampo - Finnish Culture on the Semantic Web 2.0: Thematic Perspectives for the End-user". In D. Bearman J. Trant & D. Bearman (Eds.) *Museums and the Web 2009: Proceedings*. Toronto: Archives & Museum Informatics. Published March 31, 2009. Consulted February 23, 2014. http://www.archimuse.com/mw2009/papers/hyvonen/hyvonen.html

Keller, M. A. (2011). "Linked Data: A Way Out of the Information Chaos and toward the Semantic Web". *EDUCAUSE Review*. http://www.educause.edu/ero/article/linked-data-way-out-information-chaos-and-toward-semantic-web

Keller, M. A., Persons, J., Glaser, H., & M. Calter, (2011). "Report of the Stanford Linked Data Workshop". Washington, DC: CLIR. http://www.clir.org/pubs/reports/pub152/reports/pub152/LinkedDataWorkshop.pdf

McNamee, P., Mayfield, J., Lawrie, D., Oard, D. W., & D. Doermann, (2011). "Cross-Language Entity Linking". *5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand: IJCNLP. http://www.aclweb.org/anthology/I/I11/I11-1029.pdf

Meij, E., Bron, M., Hollink, L., Huurnink, B., & M. de Rijke, (2011). "Mapping queries to the Linking Open Data cloud: A case study using DBpedia". *Web Semantics: Science, Services and Agents on the World Wide Web*, *9*(4), 418–433. doi:10.1016/j.websem.2011.04.001.

Mihalcea, R., & A. Csomai, (2007). "Wikify!: linking documents to encyclopedic knowledge". *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* CIKM'07. New York, NY: ACM. doi:10.1145/1321440.1321475.

Miller, E. and D. Wood, (2010). "Recollection: Building Communities for Distributed Curation and Data Sharing". In J. Trant and D. Bearman (Eds.) Museums and the Web 2010: Proceedings. Toronto: Archives & Museum Informatics. Published March 31, 2010, consulted February 24, 2014. http://www.archimuse.com/mw2010/papers/miller/miller.html

Milne, D. & I.H. Witten, (2008). "Learning to link with Wikipedia". *Proceedings of the 17th ACM Conference on information and knowledge management CIKM'08.* New York, NY: ACM. 509-518. doi: 10.1145/1458082.1458150.

Moretti, F. (2013). Distant Reading. New York: Verso.

NISO. "Linked Data in Libraries, Archives, and Museums". *Information Standards Quarterly* 24(2/3). Published August 1, 2012, consulted February 24, 2014. http://www.niso.org/publications/isq/2012/v24no2-3/

NIST. (2012) "Cold Start Knowledge Base Population at TAC 2012," Version 1.3, August 17. http://www.nist.gov/tac/2012/KBP/task_guidelines/Cold%20Start%202012%20Task%20Description%201.3.pdf

Paepcke, A. (2008). "An Often Ignored Collaboration Pitfall: Time Phase Agenda Mismatch". Stanford iLab blog post. http://infoblog.stanford.edu/2008/11/often-ignored-collaboration-pitfall.html

Rothberg, M. (2009). *Multidirectional Memory: Remembering the Holocaust in the Age of Decolonization*. Stanford: Stanford University Press.

Van Dijk, D. et al., (2009). "Out There : Connecting People, Places and Stories". In J. Trant and D. Bearman (Eds.) *Museums and the Web 2009: Proceedings*. Toronto: Archives & Museum Informatics. Published March 31, 2009, consulted February 24, 2014. http://www.archimuse.com/mw2009/papers/vandijk/vandijk.html

Van Hooland, S., Verborgh, R., De Wilde, M., Hercher, J., Mannens, E. & R. Van De Walle, (2013). “Evaluating the Success of Vocabulary Reconciliation for Cultural Heritage Collections”. *Journal of the American Society for Information Science and Technology*, 64 (3), 464-479.