White Paper Report

Report ID: 109308

Application Number: HD-51766-13

Project Director: Douglas Oard (oard@umd.edu)

Institution: University of Maryland, College Park

Reporting Period: 5/1/2013-4/30/2014

Report Due: 7/31/2014

Date Submitted: 8/31/2014

A White Paper on Bridging Communities of Practice: Emerging Technologies for Content-Centered Linking

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

August 31, 2014

Abstract

This white 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, librarians, archivists, museum professionals, and digital humanities scholars 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 are now contributing.

Introduction

If the three V's of "*volume*, *velocity*, and *variety*" is the mantra for "big data" today, we might coin the 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 unprecedented opportunities for access and use. In the contemporary information landscape, use may involve more than reading, browsing, or viewing; it can also involve contributing in ways that add value to materials. In particular, people can and sometimes do describe what they find in ways that they hope will have meaning to others. Digitization helps to enable this revolution in description, but digitization is just one element in a complex technical and procedural ecosystem. Another important part of this ecosystem is the ability to span distance and organizational boundaries in ways that have not previously been practical, thus offering at least the potential to describe relationships between things that are not now, and perhaps never were, physically together. Further adding to both the promise and the complexity of this new world, one person can even describe the descriptions created by another.

None of this is entirely new, of course; 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 more fully harness this 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 has come to be 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 excitement 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 white 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 libraries, archives, museums, and other institutional and individual actors 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" in which content is often found (Keller, Persons, Glaser, & Calter, 2011). 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 (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 interpretation. Moreover, since 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 a 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 interoperabilit**y**. SNAC, a collaboration of the University of Virginia, The University of California, 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 exploits 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 uncovers meaningful connections between documents and data related to the personal and professional lives of musicians, and Linking Lives (http://archiveshub.ac.uk/linkinglives/), which is an end-user interface that uses 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/; Hyvönen et al., 2009) 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 efforts.

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 materials found in LAM collections. Although computationally malleable, such user-generated content is less often drawn upon, in part because its associated metadata is not always interoperable with cultural heritage datasets, and perhaps also because many people still think of the Web more as a dissemination channel than a resource.

Machines That Learn From Us

Computer science research on 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 decade, 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 specific technologies that can directly leverage content, not just structured metadata: *wikification* (Mihalcea & Csomai, 2007; Milne & Witten, 2008) and *knowledge base population* (NIST, 2012). *Wikification* automatically builds hypertext links inside of previously unlinked content, earning the name because some of the earliest systems in this line of work learned from the way in which people have built links between Wikipedia pages. *Knowledge base population* is the more ambitious task, seeking to automatically construct linked data from unstructured content. At present, 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 content that is both easily available, which of course is not necessarily the content that will ultimately be most important to any specific users of the technology (He, de Rijke, Sevenster, van Ommering, & Qian, 2011).

The research on *knowledge base population* emerged from earlier work in the broad field of *computational linguistics* that had focused initially on information extraction, which refers to the problem of finding mentions of named entities (and other similar items) in text. 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 variants of a task sometimes referred to as *machine reading*, in which the goal was to automatically create linked data directly from the content of digital documents, could be undertaken. For the past three years, the Text Analysis Conference (McNamee, Mayfield, Lawrie, Oard, & Doermann, 2011; NIST, 2012) sponsored by the National Institutes of Standards and Technology has included a *cold-start knowledge base population task*, an ambitious effort to build “small world” models of some aspects of what is described in some coherent and moderately large (e.g., 25,000-document) collections.

In 2007, researchers in a related field, *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, Bron, Hollink, Huurnink, & de Rijke, 2011; Mihalcea & Csomai, 2007). What distinguishes *wikification* from the earlier work is the adoption of *machine learning* techniques, in which machines are programmed to learn from examples. Earlier techniques relied on the encoding of human knowledge, either directly in the structure of a computer program or encoded as rules for use by so-called *expert systems*. People and machines are both able to learn to represent and reason about common phenomena, and indeed people can often do this well based on fewer examples than *machine learning* techniques would need. But *machine learning* excels when the data volume gets so large that the human mind simply cannot remember and explain every exception to every rule. Machines excel in such circumstances, however. While each exception may be rare, aggregate exceptions to simple rules explain a substantial fraction of any “long tail” phenomenon, notably including the ways in which people use language. As a result, it is not an overstatement to claim that *machine learning* has transformed the computational manipulation of human language. Thus, machine learning lies at the heart of both *knowledge base population* and *wikification*.

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, Namata, & Getoor, 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 in structured data can be automatically constructed from less well-structured text at an unprecedented pace, the potential of graph manipulation tools that can transform that text in ways that are relevant to the needs of scholars, and the interests of the broader public that their scholarship serves, is significant.

Bringing Communities Together

Building bridges between the vibrant LODLAM community and the relevant computer science communities begins with dialogue and exchange of ideas. With that goal in mind, we organized two workshops that 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.

Workshop I

The first, two-day workshop, which we came to refer to as "LAMLink," took place in September 2013 at the University of Maryland, College Park. The two dozen participants included cultural heritage scholars and professionals, digital humanities scholars, and computer scientists. The goals of the workshop were to:

- Draw on multiple perspectives to identify important opportunities that no one community or researcher could identify alone,
- Generate ideas that would be shared with others, and
- Envision next steps to weave these communities more closely together.

Workshop sessions were organized to maximize time spent on interaction among participants. 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 produce potential employment scenarios for specific technologies, and to identify possible ways of leveraging the resulting opportunities. The workshop concluded with participants brainstorming on the design of a second workshop, ultimately recommending a full-day, hands-on pre-conference workshop at the annual conference of Code4Lib (http://code4lib.org/).

Workshop II

The second workshop, "Computational Linguistics for Libraries, Archives, and Museums" (CLLAM), took place as a pre-conference workshop at the code4lib conference in Raleigh, NC, on March 24, 2014.[1] code4lib is a community that brings together hackers, designers, architects, curators, catalogers, artists and instigators who largely work for and with libraries on applications of emerging technologies. The full workshop included about a dozen participants in addition to the organizers. As we had hoped, participants were principally practitioners, including systems librarians and software developers.

As with LAMLink, the CLLAM agenda emphasized interaction among workshop participants. Morning presentations were short, intended to showcase an array of tools and applications that could be repurposed in library and archives. Presentations emphasized how

[1] In addition to Douglas Oard, and Amalia Levi, organizers of this workshop included LAMLink participants Robert Warren (Dalhousie University) and Corey Harper (New York University). CLLAM was one out of 19 pre-conference workshops at code4lib.

technologies currently being developed by computer scientists might help to create new opportunities for cultural heritage institutions when seeking to maximize the value of large born-digital and digitized collections. These morning presentations provided a basis for rich interaction, with the afternoon devoted to a joint exploration of practical needs and the capabilities and limitations of current technologies, often in the context of specific collections and tools. Participants from a number of organizations working with very large digital collections, including the Digital Public Library of America, HathiTrust, OCLC, and the New York Public Library, brought a valuable "big data" perspective.

A Bonus Round: Engaging Museum Professionals

We were pleased to also have the opportunity to discuss what we had been learning over the course the two workshops with a paper presentation at the Museums and the Web Annual Conference in Baltimore, MD on April 3, 2014, just shortly after the CLAAM workshop. The paper we presented, an earlier draft of this white paper, focused (because of publication schedules) on lessons learned from our first (LAMLink) workshop (Oard, Levi, Punzalan, & Warren, 2014). The presentation was followed by some discussion, which helped to deepen our understanding of some of the key issues that we discuss below.

What We Have Learned

In this section, we describe our personal interpretation of the outcomes of the discussions in the two workshops and at the Museums and the Web conference.

Communities: Forging 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, just 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 has already been some cross-fertilization 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 how LOD can be valuable for humanities, but also what value the needs of the humanities can bring to computer scientists as well. Computer scientists can fairly easily measure the accuracy of the decisions their systems make, but ultimately it is the effects of those errors, not their mere existence, that scholars and LAM stakeholders care about. Moreover, computer scientists typically measure how well their tools do on data similar to what they saw during training, but the real world is of course considerably messier than that. Evaluation resources generally, and evaluation measures in particular, are thus an important 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 should be possible to make progress in this regard if LAM stakeholders and humanities scholars can to learn to help guide the work of the computer scientists in productive directions by working with them on selecting cultural heritage materials that challenge existing tools in specific ways, and on designing evaluation measures that reflect the results that are most needed (for an example, see Petras, Bogers, Ferro & Masiero, 2013).

The value of small projects: When building things, the key to success 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.

Reward structures: Over time, different communities have evolved different incentives that help to guide the work of their community. It is important to understand the degree to which these reward structures are consonant, and where new structures are needed, with an eye towards possibly instigating improvements. The recent interest in cyberinfrastructure for the humanities offers one promising direction for crafting consonant reward structures, focusing as it does on creating capabilities rather than on creating artifacts.

Technology: From Data to Knowledge

The knowledge graph: By encoding aspects of our interpretations as graphs that encode relationships between “things, not just strings,” we gain access to a powerful new mode of expression. This power arises from coupling the machine’s ability to find patterns with 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 there is benefit to bringing different communities together to see what is common across these settings, we must also bear in mind that these communities 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.

Multi-perspective LOD: Computer science, and indeed library science as well, is often inclined to view data and metadata as fixed and objective, while of course humanists also find interest in the complexities of the cultural processes that gave rise to that data. If we are to bring these worlds together, we will ultimately need platforms that support exploration in ways that are tailored to help the user identify and analyze the full complexity of the cultural constructs that LOD can reflect. At the heart of LOD is that authority derives from the source of the description rather than that which is being described, so allowing multiple concurrent and inconsistent interpretations of the content to exist, to be found, and to be further interpreted will be important.

A big technology tent: A comprehensive effort to leverage these new opportunities will need to involve an even broader range of technologists in the discussion than has been possible in this project, including, for example, experts in digital imaging and image processing, multimedia information systems, optical character recognition, and human-computer interaction.

User-centered design: Interestingly, computer science and LAM institutions both see themselves as existing to serve users, but perhaps with somewhat different users in mind. LAM institutions are one class of users for what computer scientists seek to build, but LAM institutions themselves in turn exist to serve their own users. Those ultimate users are the real people in which all of this is grounded. If we are to avoid *Field of Dreams* approaches ("if you build it, they will come"), it might help to work in teams that include humanities scholars, computer scientists, and LAM institutions that have some experience with mediating between their disparate ways of knowing.

Implications for Cultural Heritage Institutions

Co-dependence of digitization and description: Currently, LAM collections that have been indexed, cataloged, and described far outnumber collections that have been fully digitized. Moreover, the uptake of "more product, less process" approaches (Greene & Meissner, 2005) in archives is further 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, such approaches could indirectly incentivize additional digitization by providing 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 ultimately 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 cultural heritage 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), this recent development is welcome. 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 contextualize what they see by adding their own links 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 these kinds of connections easily, and technologies to support this kind of reuse have, for example, been recurring themes in *Museums and the Web* conferences (Straup Cope, 2009; van Dijk, Kerstens, & Kresin, 2009; Miller & Wood, 2010). 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). Importantly, LOD enables LAM institutions to participate in "mash-up" culture while retaining a reference to the original source.

The value of error-tolerant workflows: The dirty little secret of virtually every effort to automate the processing of human language is that the results are what might charitably called "imperfect." Indeed, "imperfect" is a bit of an understatement: in some of the most bleeding edge technologies the state of the art techniques are wrong half the time. Of course, some of what current Web search engines return is wrong too (in the sense that it is not what the searcher was looking for), but our response to that is to find problems that those search engines can help with, and then develop ways of using imperfect search engines in such cases. We might call this approach an error-tolerant workflow. If LAM institutions are to make the most of emerging technologies, we need to not only better adapt those technologies to their needs, but also to adapt the needs they seek to address and the ways they seek to address those needs to the capabilities and limitations of those new technologies. With this in mind, LAM institutions could benefit from exploring the potential for developing new error-tolerant workflows that could leverage imperfect technologies in order to provide specific types of usable services that would help advance their mission.

Next Steps

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

Institutional Structure

Bridging humanities and computer science research makes it possible to do some things that need to be done, but also some things for which we might not yet have thought through all of the consequences. To understand the implications of this intervention, we will need to educate a new generation of scholars who are adept at 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, nor 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. 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 ones we have described are useful as first steps. Ultimately, however, 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 values, you should look not just to its policy statements, but also 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 LAMLink 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 competitiveness (e.g., NSF). Any of these sources might generate further advances that we can leverage, and some 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 funding agencies to come together around joint programs of mutual interest. Here, the Digging into Data program, 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 collaboration this approach to 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. Supporting teams who work together to select and assemble such collections as boundary objects between disciplines 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 would not call it research.

A Legacy in Lieu of a Conclusion

We expect that the professional connections among participants that have been built over the course of this project will continue to pay dividends. The lasting value of what we have done together will be found not just in what we have learned in this short time, but also in what those of us who have come together for these initial exploratory discussions, and others who will build upon our work, will achieve in the coming years. We are grateful for the generous support for this process of the National Endowment for the Humanities.

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 indebted to our advisory board for their wise counsel as we tried to first bring these communities together in the LAMLink workshop, to the participants in both workshops and to those who stopped by to speak with us at *Museums and the Web* for generously sharing their insights and perspectives, to our fantastic student volunteers at the LAMLink workshop, to the College of Information Studies at the University of Maryland and the conference organizers for *Code4Lib* and *Museums and the Web*, and to our CLLAM co-organizers Corey Harper and Rob Warren. Without them, this project, and thus this white paper, simply could not have achieved what it did. Our role here is that of the messenger – we have sought to reflect what we all learned together from each other. We emphasize, however, that our reporting in this white paper 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., & Koloski, L. (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 & D. Bearman (Eds.), *Museums and the Web 2009: Proceedings*. Toronto: Archives & Museum Informatics. Retrieved February 24, 2014, from http://www.archimuse.com/mw2009/papers/cope/cope.html

Diehl, C. P., Namata, G., & Getoor, L. (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). Retrieved February 24, 2014, from http://www.digitalhumanities.org/dhq/vol/7/1/000143/000143.html

Greene, M., & Meissner, D.E. (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., & Qian, Y. (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 J. Trant & D. Bearman (Eds.), *Museums and the Web 2009: Proceedings*. Toronto: Archives & Museum Informatics. Retrieved February 23, 2014, from http://www.archimuse.com/mw2009/papers/hyvonen/hyvonen.html

Keller, M. A., Persons, J., Glaser, H., & Calter, M. (2011). *Report of the Stanford Linked Data Workshop*. Washington, DC: CLIR. Retrieved February 23, 2014, from http://www.clir.org/pubs/reports/reports/pub152/LinkedDataWorkshop.pdf

McNamee, P., Mayfield, J., Lawrie, D., Oard, D. W., & Doermann, D. (2011). Cross-Language Entity Linking. *5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand: IJCNLP. Retrieved February 23, 2014, from http://www.aclweb.org/anthology/I/I11/I11-1029.pdf

Meij, E., Bron, M., Hollink, L., Huurnink, B., & de Rijke, M. (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., & Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* CIKM'07. doi:10.1145/1321440.1321475.

Miller, E., & Wood, D. (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. Retrieved February 24, 2014, from http://www.archimuse.com/mw2010/papers/miller/miller.html

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

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

National Information Standards Organization [NISO]. (2012). Linked Data in Libraries, Archives, and Museums. *Information Standards Quarterly 24*(2/3). Retrieved February 24, 2014, from http://www.niso.org/publications/isq/2012/v24no2-3/

National Institute of Standards and Technology [NIST]. (2012). *Cold Start Knowledge Base Population at TAC 2012* (Version 1.3, August 17). Retrieved February 24, 2014, from http://www.nist.gov/tac/2012/KBP/task_guidelines/Cold%20Start%202012%20Task%20Description%201.3.pdf

Oard, D., Levi, A., Punzalan, R., & Warren, R. Bridging Communities of Practice: Emerging Technologies for Content-Centered Linking. In N. Proctor & R. Cherry (Eds.), *Museums and the Web 2014: Proceedings*. Silver Spring, MD: Museums and the Web. Retrieved August 28, 2014, from http://mw2014.museumsandtheweb.com/paper/bridging-communities-of-practice-emerging-technologies-for-content-centered-linking/

Paepcke, A. (2008, November 8). An Often Ignored Collaboration Pitfall: Time Phase Agenda Mismatch". *Stanford iLab blog*. Retrieved August 28, 2014, from http://infoblog.stanford.edu/2008/11/often-ignored-collaboration-pitfall.html

Petras, V., Bogers, T., Ferro, N., & Masiero, I. (2013). Cultural Heritage in CLEF (CHiC) 2013 – Multilingual Task Overview. In *CLEF 2013 Evaluation Labs and Workshop Online Working Notes*. Retrieved August 28, 2014, from http://www.clef-initiative.eu/documents/71612/82b4444b-9a3c-4d8e-a986-6a184012991e

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

Van Dijk, D., Kerstens, K., & Kresin, F. (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. Retrieved February 24, 2014, from http://www.archimuse.com/mw2009/papers/vandijk/vandijk.html

Van Hooland, S., Verborgh, R., De Wilde, M., Hercher, J., Mannens, E., & Van De Walle, R. (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.