IFLA WLIC 2018
KUALA LUMPUR

Submitted on: 27.07.2018

'Digging Deeper, Reaching Farther: Libraries Empowering Users to Mine the HathiTrust Digital Library': An Overview

Harriett Green
University Library, University of Illinois at Urbana-Champaign, Urbana, U.S.A.
E-mail address: green19@illinois.edu

Abstract:

The project "Digging Deeper, Reaching Further: Libraries Empowering Users to Mine the HathiTrust Digital Library" is a three-year initiative funded by the Institute for Museum and Library Services that aims to prepare librarians for the new, data-driven research landscape by developing and disseminating a curriculum specifically designed to teach digital scholarship competencies to librarians and LIS professionals, with a focus on computational text analysis. The project developed a professional development that builds librarians' text data mining skills by using the HathiTrust Research Center (HTRC) as the anchoring example resource for pursuing text analysis research. The project staff launched this curriculum through a series of workshops held throughout the United States, and it will be released as an open educational resource at the end of the grant in 2018. This paper will detail the implemented strategies to develop the 'train the trainer' curriculum, and conduct workshop events that taught librarians about text data mining, and concludes with early findings from the project's research investigation on how librarians seek to support digital scholarship at their home institutions.

Keywords: Digital scholarship, professional development, data science, digital libraries.

Introduction

As trends in digital scholarship spread across the disciplines, librarians are called to play an increasingly active role in digital scholarly methods and research partnerships with faculty. Librarians have long been embedded in the research process, and with the advent of digital humanities, computational social sciences, and other data-driven research approaches,

librarians are seeking to learn the digital tools and computer programming skills that undergird these emergent areas. Yet many librarians did not receive training in computer programming and digital methods in their LIS education, and the opportunities to gain such computational skills are often a challenge to find and pursue. How do we prepare those currently in the LIS profession for the new digitally-oriented research occurring today? The "Digging Deeper, Reaching Further" project (hereafter referred to as "DDRF") seeks to address this professional development need by developing a "train the trainer" curriculum and program focused on digital scholarship skills for librarians, with a specific focus on text analysis. The project received a three-year Laura Bush 21st Century Librarian grant award from the Institute of Museum and Library Services (award #RE-00-15-0112-15), and began its work in fall 2015. The project is led by the University of Illinois at Urbana-Champaign in partnership with Indiana University Bloomington, Lafayette College, Northwestern University, and the University of North Carolina at Chapel Hill, and together this collaboration represents diverse institutions and librarian communities.

Background

Training opportunities for digital humanities and digital scholarship have grown over the past decade, and include such organized programs as the Digital Humanities Summer Institute (https://dhsi.org), the Humanities Intensive Learning and Teaching (HILT) institute (http://www.dhtraining.org/hilt/), and DH at Oxford (http://www.dhoxss.net/) as well as other programs in the DH Training Network, which equip attendees to engage in digitally-intensive research. For librarians, digital scholarship-related professional development opportunities have emerged only in recent years, including the Library Carpentry curriculum and workshop series (Baker et al., 2016), Digital Humanities Institute for Mid-Career Librarians at the University of Rochester (http://humanities.lib.rochester.edu/institute/), the Data Science and Visualization Institute for Librarians at North Carolina State University (https://www.lib.ncsu.edu/datavizinstitute), and the Association of Research Libraries' Digital Scholarship Institute (Melton et al. 2015). But these programs require resources that not all librarians and scholars can access, and while there is a proliferation of self-guided resources such as CodeAcademy and Data Camp, they leave aspects of librarianship and research collaborations unaddressed.

Studies document the growth in uniquely intensive research collaborations between librarians and disciplinary researchers (Green, 2014; Auckland, 2010), and there is now a diversity of case studies on how librarians partner on digital humanities research (Hartsell-Gundy, Braunstein and Golomb 2015; Gilbert and White, 2016). Faculty research partnerships (Alexander et al., 2014), research tool development (Nowviskie 2014), scholarly communication engagement (Coble, Potvin, and Shirazi 2014), and research centers featuring multiple services and resources for digital scholarship research support (Vinopal and McCormick 2013) are becoming increasingly standard in libraries.

In this context, the DDRF project aims to empower librarians--especially those without local training programs--to become active in digital scholarship on their campuses. As such, DDRF seeks to build capacity in support of the Institute for Museum and Library Services (IMLS) National Digital Platform initiative.

Workshop Design and Development

Curriculum Development

Curriculum development was the focus of the first six months of the project, and a sub-set of project team members from Illinois and Indiana formed the Curriculum Working Group to

write and iterate the curriculum materials. Two particularly important insights gained from this process of creating the initial curriculum content were the importance of strategic and thoughtful lesson planning, and the value of consulting external experts.

As part of this planning process, the project staff conducted informational interviews with organizers of other librarian training initiatives, including the University of Rochester's Digital Humanities Institute for Mid-Career Librarians and North Carolina State University's Data and Visualization Institute for Librarians. These fruitful conversations provided the team with useful insights and lessons learned that informed the design of the curricular materials, and also helped us forge key relationships with these related initiatives.

The Working Group then considered and discussed the needs of the librarian learner audience, after which they outlined a set of coherent and practical learning goals and outcomes. These learning goals are tied to specific competencies and hands-on activities that build in technical skills using the scaffolding approach for learners at different levels. The learning goals and outcomes address librarian-specific competencies to engage with digital scholarship, and they were guided by the idea of offering text data mining as a digital scholarship service supported by the library. The learning outcomes do not aim for the learner to become an expert over the course of several hours, nor for the learner to necessarily formulate their own research project. Instead, the learning goals focus on fostering awareness of, and the ability to communicate about, key tools and methods in text analysis.

The curriculum is anchored in the tools and services provided by the HathiTrust Research Center, and its parent organization, HathiTrust, as a lens for exploring methods and techniques for large-scale text analysis that moves beyond volume-by-volume access. The HathiTrust Digital Library contains over 16 million volumes of digitized text contributed by research libraries from around the world. Its size affords scholars the opportunity to increase the scale of their inquiry and to ask new kinds of research questions. The HathiTrust Research Center has developed a suite of tools and services for performing computational text analysis on material in the HathiTrust Digital Library (Downie et al. 2015). As such, it provided a useful anchor for teaching about text analysis approaches and techniques, from building a research corpus to analyzing it with computational methods.

From the learning goals, the team created five training modules that frame a basic research workflow for text analysis, from finding textual data to cleaning the data and analyzing it, which also align with key points at which a librarian might be involved in the research process (Table 1). Each module incorporates skills-based competencies that are developed through hands-on activities. A sample reference question that could be resolved using various approaches to text analysis is used through the modules and guides hands-on activities and discussion. Several of the modules and their activities focus on HathiTrust Research Center data, tools, and services for text mining, but the curriculum places equal attention and weight to broad, core concepts for text analysis and computational methods and tools for data mining.

The modular format for the curriculum enables librarians to adapt the materials into workshops for different settings and audiences. After soliciting feedback on the teaching materials from all the partners, the first iteration of the curriculum was ready to be piloted.

Table 1: Curriculum Modules and Content.

Module	Primary learning goal	Skills developed
Introduction	Understand what text analysis is and how scholars are using it in their research.	Recognize research questions that may lend themselves to text analysis methods.
Gathering Textual Data	Differentiate the various ways textual data can be acquired and evaluate textual data providers.	Build a textual dataset and run a web scraping script.
Working with Textual Data	Distinguish cleaning and/or manipulating data as a part of the text analysis workflow.	Clean text data files using a Python script and/or OpenRefine.
Analyzing Textual Data	Recognize the advantages and constraints of web-based text analysis tools and programming solutions.	Run a web-based text analysis algorithm and extract token frequencies from a dataset.
Visualizing Textual Data	Identify data visualization as a component of data-driven analysis.	Practice exploratory data analysis using different tools for visualization.

Design and Piloting of Workshops

The initial, modular workshop curriculum was first tested in spring 2016 at two pilot workshops at the University of Illinois and Indiana University libraries. In addition to these preliminary workshops, the iterative instructional design process included strategic adjustment of learning goals, assessment of the curriculum based on feedback from workshop attendees and project partners, and the creation of second-draft teaching materials, including slides, teaching guides, and sample datasets.

After the first pilot workshops, the Curriculum Working Group collaborated throughout the summer to adapt and refine the draft curriculum for fall 2016 pilot workshops at each of the project institutions. The overall content of the workshops remained fairly consistent across institutions, but the team made slight modifications to customize the teaching materials for each partner, and project partners provided more constructive feedback during this process.

Project team members taught workshops in fall 2016 at the five partner institutions: The teaching activities included experimenting with different workshop formats, including shorter and longer sessions, as well as splitting the content over two sessions on different days. The Curriculum Working Group revised the curriculum based on these pilot workshops, and then team members from all five partner institutions taught another round of pilot workshops in spring 2017. The feedback and experiences from these varied workshop events in different library sizes and contexts were valuable in adjusting the curriculum to the diverse audiences anticipated for the national series of workshops.

National Workshop Training Series

In March 2017, the inaugural DDRF national "train the trainer" workshop event was held at the Johns Hopkins University Libraries to coincide with the Association of College and Research Libraries (ACRL) 2017 conference in Baltimore, Maryland. This early national workshop event allowed project members to gain experience in planning and promoting learning events for librarians. Lessons learned included the value of a pre-workshop checklist for hosting institutions and of a coordinated communications program, and how to prepare the workshop schedule for a national audience.

The subsequent national series of training workshops was then the core focus for the third year of the DDRF project. Beginning in October 2017 and running through August 2018, workshops have been held in seventeen more locations in multiple regions of the United States, ranging from Los Angeles, Seattle, Honolulu, Denver, and San Diego in the West; Houston, New Orleans, and Atlanta in the southeastern U.S.; Chapel Hill (North Carolina), Washington D.C., Boston and New York in the mid-Atlantic and northeast U.S.; and Chicago, Minneapolis, Lawrence (Kansas), and Milwaukee in the Midwest, among other locations. For most of the events, the project team was able to partner with major academic libraries in the regions to host the workshops, with the host institutions generously providing facilities, A/V equipment, and even refreshments, while the grant covered all travel expenses for the two to four instructors who co-taught at each event and all workshop materials. Over 300 professionals have attended these workshops, and the participants have opportunity to continue engaging via the virtual forum hosted with the Commons in a Box software.

Workshop Findings

The team learned several lessons from the pilot workshops: First, for workshops that are highly dependent on technology and delivered in computer labs, it is crucial to be familiar with the policies of the labs and test out any equipment beforehand whenever possible. Based on our experience, the team decided to use web-based systems as much as possible to mitigate installation and platform issues, though it required ensuring that there are robust wireless connections available. Second, it is important to balance the content, flow, and pace of workshop delivery. For example, after the pilot workshops, we realized that they had overwhelmed attendees by listing the names of the many tools they could use. We thus adjusted the in-person workshop content to showcase representative examples, and provide post-workshop resources with more comprehensive information.

After each workshop in both the pilot rounds and national training series, participants provided responses to a feedback form and assessment survey. From these assessment survey responses, we continually learned ways to improve the curriculum's effectiveness. that attendees desired as much hands-on content as possible. They were mixed in their feedback of how technical they wanted the workshop to be, with some wishing that the workshop covered programming concepts and skills in more depth than they experienced, while others thought we spent too much time on the "nuts and bolts." Attendees reported wanting access to the workshop materials before and after the session, and they suggested a follow-up workshop or dividing the content into true beginner and advanced session.

In summer 2017, the project team solicited feedback from the four-person DDRF Advisory Board, and they provided in-depth suggestions to improve the instructional design of the workshops. Additionally, project staff from each of the partner institutions met in Chicago in June 2017 to do a thorough, face-to-face review of the curricular materials in a full-day

retreat. Project team members reported that this event was beneficial for them to meet their collaborators in-person and spend focused time on the curriculum together. The recommendations from the Advisory Board and outcomes from the retreat were incorporated into the curriculum.

The project team updated the curricular materials throughout the series of pilot workshops, with an initial set of revisions in December 2016 and January 2017, and a second iteration of curriculum revisions in summer 2017. In response to the assessment from the second round of pilot workshops, they incorporated more concrete examples into the workshops and adjusted the pacing, removing content and providing more discussion to allow for a less rushed and more reflective learning experience. In the extensive revisions carried out in summer 2017 in preparation for the national series, the team streamlined the lectures, augmented the lectures with more participatory discussions and concepts, and refined the exercises with the HTRC algorithms and Python scripts. In summer 2018, the project team continued to make minor modifications to the curriculum, as well as began to prepare the teaching materials to be released as an Open Education Resource.

Early Assessment Findings

Through regular, systematic assessment of the curriculum and training program, the project team not only gathers feedback that informs the preparation of the final release of the curriculum materials as an Open Education Resource, but also conducts a research investigation on professional development needs for LIS practitioners in digital scholarship. Through a comprehensive assessment survey and semi-structured interviews with workshop attendees, the project aims to examine effective instructional design and pedagogical strategies for training LIS practitioners in text analysis research methods, and how to support librarians in building skills and strategies for supporting digital scholarship research at their home institutions.

After each workshop, participants voluntarily complete assessment surveys that ask them to evaluate various aspects of the workshop curriculum, pedagogical methods, and structure, and they also share their perspectives on learning needs for librarians in the area of digital scholarship. As of this writing, survey responses are still being gathered from the final training workshops in August 2018, and quantitative and qualitative analysis of the survey responses is ongoing. But early basic analysis of assessment survey responses received from attendees at the first six national workshops has begun to reveal certain aspects of LIS professional needs for digital scholarship.

In early analysis of the survey results gathered through February 2018, we reviewed 72 responses that were gathered from attendees the six national training workshops held from October 2017 through February 2018 at the University of Minnesota, Northwestern University, Emory University, University of North Carolina-Chapel Hill, UCLA, and University of Denver. Participation in the assessment surveys were wholly voluntary.

Most identified as themselves as subject specialists or liaison librarians (34.29 percent), with about 14.2 percent as reference librarians and about 12.9 percent identifying as digital scholarship/digital humanities librarians. Nearly half of the participants had been in the library field for 3 to 10 years, and about 20 percent for more than 20 years. About a third of the participants had not attended other similar/related trainings in the last 2 years (or did not

answer the question), while others had attended trainings at venues such as the Digital Humanities Summer Institute, the DLF Forum, Data Carpentry, DH+Design, or the Data Science and Visualization Institute.

Responses to the questions about the workshop content and their experiences revealed rising levels of familiarity with the tools: When asked about their comfort level with three of the main tools used in the curriculum—the command line interface, Python programming, and web-based text analysis tools—respondents generally expressed an improvement in their comfort level on a seven-point Likert scale (Extremely comfortable, Moderately comfortable, Slightly comfortable, Neither comfortable nor uncomfortable, Slightly uncomfortable, Moderately uncomfortable, or Extremely uncomfortable). For instance, 40 participants (55.6 percent) indicated a rise in their comfort level with the command line interface after the workshop, and 6 out of the 10 participants who had felt extremely uncomfortable with the command line interface prior the workshop then selected slightly comfortable after the workshop, a difference of four points on the scale.

Figure 1: Comfort level with command line interface before workshop.

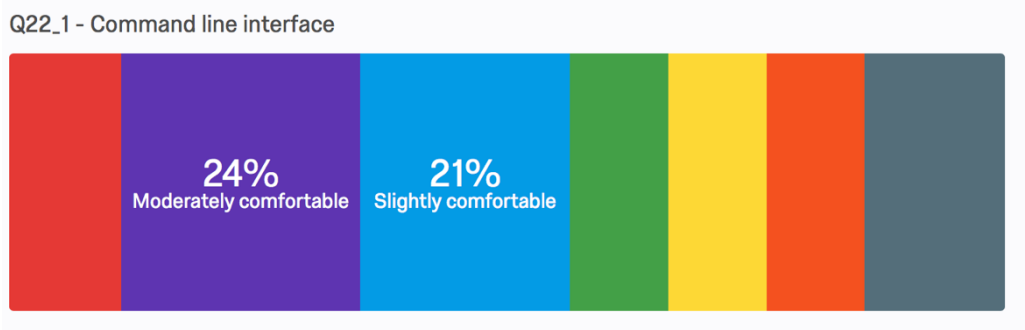
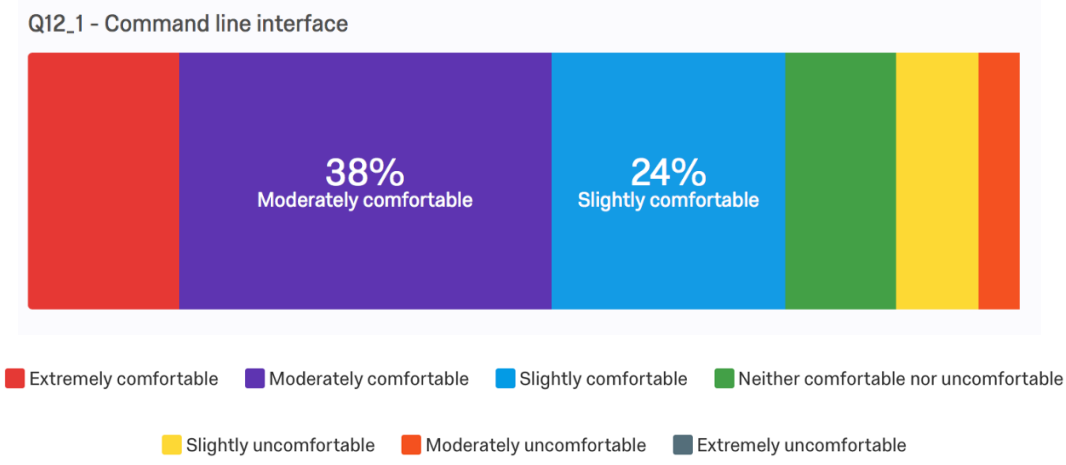


Figure 2: Comfort level with command line interface after workshop.



The participants also were asked to rate different aspects of the workshop, including the materials, the activities, the teaching methods of the instructors, and the overall organization of the workshop, and the workshop materials rated the highest and the other aspects receiving high to moderate effectiveness (Figure 3).

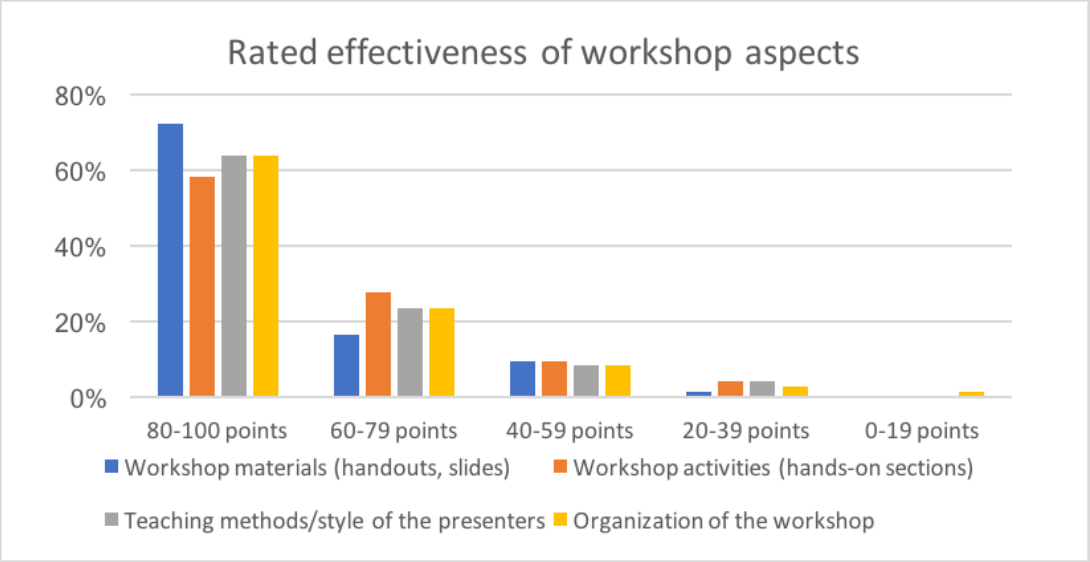
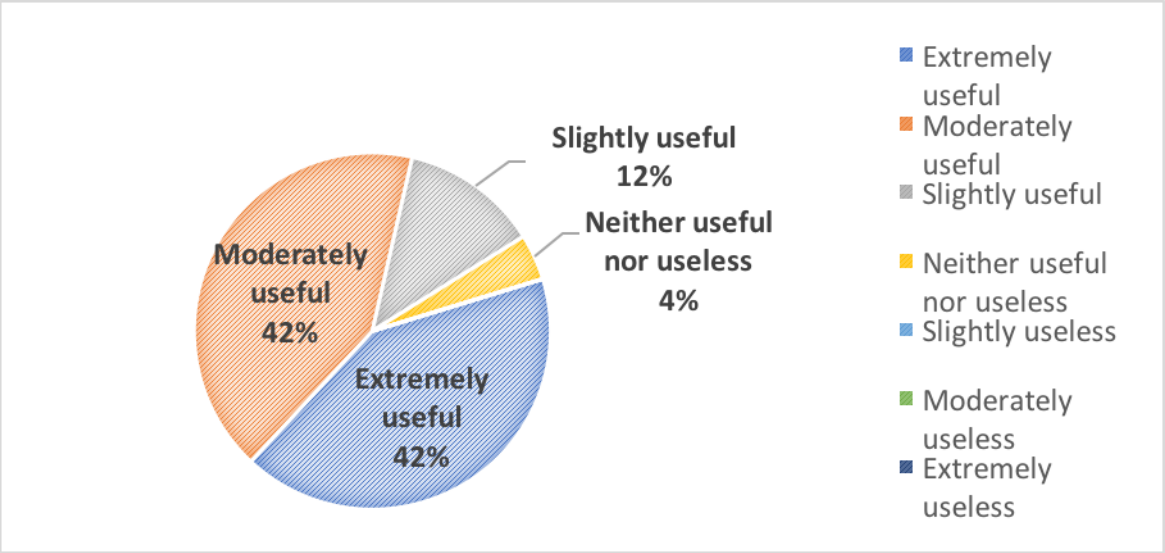


Figure 3: Rated effectiveness of workshop aspects.

In terms of the overall usefulness of the workshop for their professional development, approximately 84 percent found the workshop to be Moderately Useful or Extremely Useful (Figure 4).



When asked for what they wished could be improved or added, participants expressed the desire for more hands-on aspects and having more time for working on project, with one participant suggesting "It would be cool if there could be a second day of this workshop where we get to work on our own projects, or work through a simulated project. I understand each step individually but would like a supervised chance to put it into practice as a whole." Other attendees shared the very positive benefits they received from the event, with one respondent saying, "Thank you! There was a tremendous amount that we got through, and I left feeling happy and smarter," and another noting, "I loved this workshop! Actually working in PythonEverywhere was so empowering!"

When asked about what kinds of skills that librarians needed to acquire to engage in digital scholarship research, over 20 percent mentioned programming skills, including Python and R. Several others also mentioned acquiring additional knowledge and/or practice on digital humanities/text analysis in general, getting more hands-on experience with working on datasets, and some felt that they needed more knowledge and tools for finding related resources and sources to acquire data. Other specific areas and tools where librarians might need increased skills included statistics, data cleaning, data visualization, metadata, project management, API protocols, GIS, and Excel.

This preliminary analysis of the initially gathered survey data begins to reveal key themes in what academic librarians view as the professional development skills needed to engage with digital scholarship and data science, and effective pedagogical approaches for inculcating skills in digital research methods. The insights from these survey responses will be further augmented by semi-structured interviews with attendees: At the end of the survey, respondents are asked if they are willing to participate in an in-depth interview, and if so, they are taken to a separate form where they provide their contact information. Approximately 25 interviews have been conducted to date, and the project team is just beginning to review and conduct qualitative coding of the interview transcripts.

As we continue to analyze the gathered data, we hope to reveal insights into how the profession can further build effective professional development for librarians in digital research methods and how librarians can begin to more deeply engage in research collaborations.

Conclusion

The enthusiastic participation and responses to the training workshops and curriculum materials produced by the DDRF project over the past two years has evidenced that LIS professionals are actively seeking sustainable and accessible training in data science and digital scholarship. Through this re-skilling of library and information professionals, the aim of the DDRF project, especially within the context of the IMLS's National Digital Platform initiative supporting it, is to situate the academic library as a learning space for encountering the "big data" tools and methodologies that are being made more broadly accessible through the ever expanding suite of digital tools and open data: By equipping and empowering librarians to learn and then teach digital research methods through the lens of text data mining via an accessible curricula, the DDRF project aims to increase access to an important digital service through instructional interventions that will ultimately improve users' experience in conducting cutting-edge explorations in a unique research environment.

Acknowledgments

Thank you to the Institute for Museum and Library Services for their funding support of this project (award #RE-00-15-0112-15). Thank you to the DDRF Advisory Board members Char Booth, Francesca Gianetti, Miriam Posner, and Claire Stewart. Many, many thanks to Eleanor Dickson, Ruohua Han, and Leanne Nay on the Curriculum Working Group for their immense investment of time, energy, and careful thought in developing the DDRF curriculum and materials. Thank you to Eleanor Dickson, Ruohua Han, Rachel Blomer, and the rest of the project's Assessment Working Group for their initial analyses of the data. And a big thanks to all of the DDRF project team members over the past three years at the University of Illinois, Indiana University in Bloomington, Northwestern University, University of North

Carolina-Chapel Hill and Lafayette College for their excellent collaborative work in developing the curricular materials, organizing and leading workshops, conducting research, and overall contributing significant efforts to make this project come to successful fruition.

References

Auckland, M. (2012). *Re-skilling for research: An investigation into the role and skills of subject and liaison librarians required to effectively support the evolving information needs of researchers*. London: Research Libraries UK. Retrieved from http://www.rluk.ac.uk/wp-content/uploads/2014/02/RLUK-Re-skilling.pdf

Alexander, L., Case, B., Downing, K., Gomis, M. & Maslowski, E. (2014). Librarians and scholars: Partners in digital humanities. *EduCause Review,* http://er.educause.edu/articles/2014/6/librarians-and-scholars-partners-in-digital-humanities;

Baker, J. et al., (2016). Library Carpentry: software skills training for library professionals. *LIBER Quarterly*. 26(3), 141-162. DOI: http://doi.org/10.18352/lq.10176

Coble, Z., Potvin, S., & Shirazi, R. (2014). Process as product: Scholarly communication experiments in digital humanities. *Journal of Librarianship and Scholarly Communication* 2(3), eP1137. http://dx.doi.org/10.7710/2162-3309.1137

Downie, J.S., Furlough, M., McDonald, R.H., Namachchivaya, B., Plale, B.A., & Unsworth, J. (2016). The HathiTrust Research Center: Exploring the full-text frontier. *EduCause Review*, May 2, 2016, http://er.educause.edu/articles/2016/5/the-hathitrust-research-center-exploring-the-full-text-frontier

Gilbert, H. & White, J. eds. (2016). *Laying the foundation: Digital humanities in academic libraries*. Lafayette, IN: Purdue University Press.

Hartsell-Gundy, A., Braunstein, L. & Golomb, L. eds. (2015). *Digital humanities in the library: Challenges and opportunities for subject specialists*. Chicago: Association for College and Research Libraries.

Melton, S., Dalmau, M., Dimmock, N., Tracy, D., & Glass, E. (2017). "ARL Digital Scholarship Institute." *Digital Humanities 2017 Conference Abstracts, McGill University and Université de Montréal, Montréal, Canada, August 8-11, 2017.* https://dh2017.adho.org/abstracts/112/112.pdf

Nowviskie, B. (2013). Skunks in the library: A path to production for scholarly R&D. *Journal of Library Administration* 53(1), 53-66 http://dx.doi.org/10.1080/01930826.2013.756698

Vinopal, J. & McCormick, M. (2013). Supporting digital scholarship in research libraries: Scalability and sustainability. *Journal of Library Administration* 53(1), 27–42.