Project Abstract

The University of Pittsburgh seeks a National Forum Grant to build institutional capacity under the Laura Bush 21st Century Librarian Program. As librarians increasingly use, build, and maintain the National Digital Platform, the skills to manipulate, analyze, and manage data are crucial. Librarians will need to leverage the tools and techniques of data science, but they currently face two significant challenges:

1. *The Skills Gap* - While practicing mid-career librarians are learning some data science skills, it is through ad-hoc, uncoordinated continuing education programs.
2. *The Management Gap* - Library administrators need toolkits and frameworks to strategically leverage data science for data-driven decision making and management of library operations.

Our goal is to coalesce diverse and disparate communities whose experiences and perspectives on data science inspire and contribute to *developing* and *sustaining* our National Digital Platform. We are at a critical moment where a plethora of opportunities and promises around data science need to be *thoughtfully* and *strategically* applied in libraries. Our objectives are to:

- Articulate a vision for data science in libraries balancing data's transformative potential with a critical awareness of privacy, equality of access, and ethics in data-rich environments.
- Develop a roadmap with clear pathways for developing the next generation of continuing education programs focused on data science skills for librarians.
- Provide guidance and recommendations for preparing the next generation of library administrators on managing data intensive teams and organizations.

This proposal builds upon experiences running Data Science Training for Librarians(DST4L) a program started in 2012 at the Harvard/Smithsonian Center for Astrophysics. DST4L is one of many similar programs whose aim is to address the technical *skills gap* of mid-career librarians. But lack of coordination across these programs has resulted in an uneven landscape of orientation and objectives. This community needs an opportunity to come together to share their experiences and develop a collaborative vision.

Our experiences have shown while attendees of continuing education programs express enthusiasm in learning technical skills, when they return to their home institutions they cannot effectively apply their newly acquired skills in the course of their work. Without the appropriate organizational and technical infrastructures in place, a practicing librarian's expertise cannot reach across *the management gap.* Conversely, Library administrators are not benefitting from data science capability and libraries are not realising the full potential and power of data science through the cultivation of new work practices.

We address the skills and management gaps by bringing together diverse constituencies to discuss opportunities for coordination and collaboration. Overcoming these gaps is crucial for connecting and interconnecting the National Digital Platform to the other domains already transformed by data science. We will convene a multi-day consultation workshop at the University of Pittsburgh. The focus of the workshop will revolve around the following questions:

1. What are the technical skills and roles needed to effectively leverage data science in libraries? Is there a single set of skills or an ecology of expertise?
2. How can the existing community continuing education programs work in concert to more effectively share resources, expertise, and technical infrastructure? How tightly or loosely should they coordinate and what communities are currently being left out?
3. What can these programs do to address the management gap and support data savvy librarians beyond individual workshops?
4. What is a compelling vision for data science librarianship that activates library administrators towards effectively and critically leveraging data science for decision making?

The Data Scientist as the 21st Century Librarian?

1. Statement of Need

Advances in statistics and computer science, combined with an abundance of data from digital infrastructure, web-based user transactions and online services, have given rise to a new professional ecosystem called data science. Data science has many definitions, but at its core it is about "*generating insight from data to inform decision making*."[1] Data science methods and products have transformed commerce, healthcare, and government and they will continue to transform other sectors. *The Federal Big Data Research and Development Strategic Plan,* recently released by the Obama Administration Big Data and Research Initiative, explicitly identifies curators, librarians, and archivists as *core specialists* to help to meet growing demand for analytical talent and capacity across all sectors of the national workforce. As society is increasingly infused with data, librarians will have a crucial role in the future development of the data science ecosystem across multiple sectors; the report acknowledges "*investments are needed to expand the current pipeline of support to the field of data science*."[2]

In 2015, the ALA published a special issue of *Library Technology Reports* by librarian and software developer Andromeda Yelton, illustrating the practical benefits of librarians with technical skills. By acquiring programming skills, they were more effective at organizing information guides, generating automated reports, and processing cataloging records. While some iSchools and Library and Information Science programs have started adding technical requirements to their programs, these requirements are not evenly distributed or part of core curricula (Burton 2016). Furthermore, such graduate training does not help mid-career librarians, who want and need to cultivate new technical expertise in order to better serve their patrons. Librarians who are familiar with the life-cycle of data-intensive research data collection, cleaning, visualization, and analytics, can better support and acquire credibility with research faculty in the sciences, social sciences, and humanities.

Universities are making significant investments in data science across their campuses[3] and librarians will need to be prepared to contribute to these efforts. The ACRL 2015 Environmental Scan[4] identified a need for more advanced data curation services, urging librarians to have a deeper knowledge of domain research practices to enable them to help researchers with data management, sharing, and preservation. Librarians need both generic cross-disciplinary skills and specific subject knowledge (Mayernik et al. 2014). A recent ALL-SIS report[5] identified the need for librarians to have technical data curation expertise, especially in the areas of data visualization and text mining. For example, text mining skills would be advantageous for law librarians working with legal scholars, and in particular when combined with an understanding of intellectual property and licensing issues surrounding access to legal databases. It is precisely this kind of *blended* domain knowledge, technical skills and librarians' informatics expertise representing capability as a data scientist, which will have a transformational impact on professional roles, associated practices and the perceived value of libraries more widely.

[1] http://web.archive.org/web/20160304151135/http://christianlauersen.net/2016/01/11/librarians-as-data-scientists/
[2] https://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/bigdatardstrategicplan-nitrd_final-051916.pdf
[3] http://msdse.org/environments/
[4] http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/EnvironmentalScan15.pdf
[5] http://www.aallnet.org/sections/all/storage/committees/scholarlycommunication/ALLSISscholcomm2013625.pdf

There are many continuing education programs, training camps, and informal workshops oriented towards teaching librarians technical skills. Members of the project team created data science training for librarians (DST4L)[6] out of an increasingly urgent need for technical expertise at the Harvard/Smithsonian Center for Astrophysics. "*The main objectives [of DST4L] are for participants to learn to extract, analyze and present data using the most-up-to-date techniques...our staff should have the same skills as the scientists and researchers who patronize their libraries, so they can understand their data needs better and build services that respond better to their needs*."[7] North Carolina State University has launched the Data and Visualization Institute for Librarians,[8] a week-long course to help librarians "*develop knowledge, skills, and confidence to communicate effectively with faculty and student researchers about their data*." The program focuses on deep technical skills such as data analysis, visualization, and sharing, and also deals with advanced topics such as statistical analysis or version control with GitHub. Software Carpentry[9], a volunteer-led continuing education program, has been teaching domain scientists and engineers the computing skills necessary for data-intensive research. The program emerged to fill an ever growing gap between the technical skills needed for research and disciplinary teaching. Librarians have recognized the role they can play in hosting and participating in Software Carpentry workshops to better serve their constituencies.[10] Variations, such as Library Carpentry, experiment with the Software Carpentry two-day workshop model to address the skills gap for practicing librarians, but they need stronger institutional support.[11]

The effectiveness of ad-hoc, informal, or short-term training and education programs is limited because they operate outside of traditional institutional support structures, professional development, and incentive systems. The incentive structures for mid-career librarians can be mis-aligned or opposed to the development of technical skills. Investing in professional development, like creating opportunities for librarians to attend the NSCU Data and Visualization Institute, or hosting a Software/Library Carpentry workshop, has the potential for significant benefits, but only when administrators create opportunities for the application of the newly acquired skills. Practicing librarians may be blocked by regressive organizational structures and limitations on personnel; "*I learn new skills, but I still need to do my old job.*" Job descriptions can be inflexible, making technical professional development difficult, which may discourage the acquisition of new technical skills. Organizational and managerial support is crucial if technical skills are to be acquired, effectively applied, and have an impact: "*librarians with tech-savvy managers had an easier time getting support for their coding activities (whether formal, like courses, or informal, like time to code at the office as long as the work got done)*" (Yelton 2015 p.23).

Despite the wealth of training programs dedicated to librarians acquiring data skills, there are few focused on cultivating tech-savvy managers or on operationally managing data-intensive teams.

[6] http://altbibl.io/dst4l/

[7] http://web.archive.org/web/20160318110233/http://library.harvard.edu/02042014-1336/harvard-library-offers-data-scientist-training

[8] https://www.lib.ncsu.edu/datavizinstitute

[9] http://software-carpentry.org/about/

[10] http://web.archive.org/web/20160312183652/http://software-carpentry.org/blog/2014/08/bootcamps-for-librarians.html

[11] http://librarycarpentry.github.io/fin/

Current management programs, e.g. Harvard Leadership Institute for Academic Librarians[12] provide mechanisms for developing future senior-level managers, but could provide a more specific emphasis on managing data-intensive librarians. Managers need to be aware of new data science roles (Lyon & Brenner 2015) and the specific requirements for these positions (Lyon & Mattern 2016). They also need to understand how to integrate them into their organizations and envision the diverse contexts, opportunities, and benefits in applying data science methods. Library managers and administrators need supporting frameworks and toolkits to leverage data science capability in their strategic planning and decision-making, in the cost-effective operational management of library services, and in developing librarian teams supporting data-intensive communities on campus and beyond. For many senior librarians, managers, and administrators, the value proposition of data science skills is not readily apparent. Stories of success with clear statements of impact, backed with evidential data, need to be shared as case studies. The methods and techniques of data science should be used to validate, (with rigorous documentary evidence), that librarians with data science skills and positioned within organizations (re-)configured to support them, have a clear, demonstrable value both within the library and beyond. Management programs like Library Leadership in a Digital Age[13] would be an ideal partner whose curriculum could be informed by discussions about management of data-intensive organizations.

The overarching goal of this project is to bring the target audiences of library practitioners, educators, managers / administrators, and data science communities into conversation, to foster a multi-directional flow of information, knowledge, and collaborative opportunities. More specifically, in our objectives for libraries, we aim 1) to understand the workforce skills requirements and management gaps at strategic and operational levels, and 2) to develop an ‘Implementing Data Science Roadmap’ for overcoming them. For data science, we will bring librarians into broader conversations 3) to build awareness and knowledge of data privacy, ethics and access, and 4) to identify operational contexts both internal and external to the library, where data science methods may be most-effectively applied. Finally, we aim 5) to provide a National Forum to offer co-ordination and facilitation to catalyse audience collaboration, informed discussion, agenda-setting, and will act as a springboard to bring the disparate data science stakeholders together to work on a plan to transform the landscapes of libraries and data science.

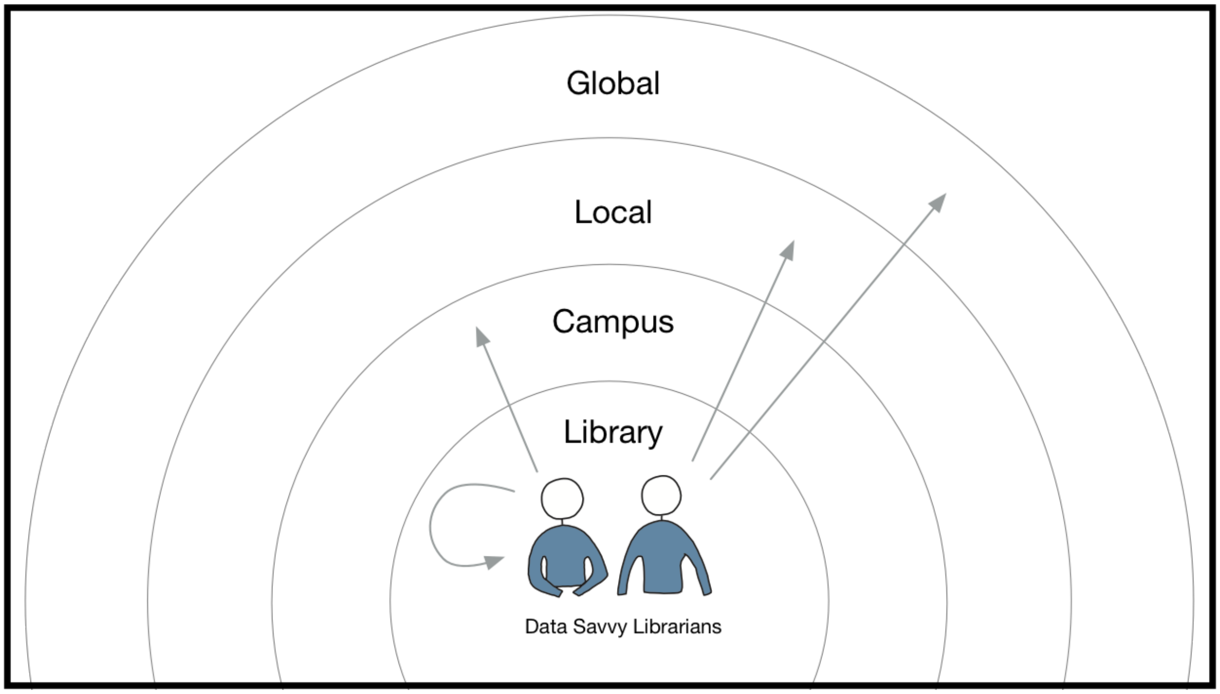
2. Impact

A National Forum is critical at this time because while many individuals, groups, and institutions are committing resources towards addressing data science in libraries, these efforts are uncoordinated and fragmented. The National Forum framed as a workshop, will bring the key constituencies together (continuing educators, library administrators, practicing librarians, industrial data scientists and other relevant parties), to address the skills and management gaps and to share the solutions they have developed. Through the workshop, participants from multiple institutions will foster personal relationships and the connections necessary to establish a community of interest around data science in libraries. They will also hear real-world case studies of successful and *innovative* implementations of data science in libraries and will understand the specific *skills* required for these new roles. Following the workshop they will be able to *transfer* their enhanced knowledge and understanding of data science, back to workplace colleagues within their organizations. Subsequent follow-up actions will identify the degree of *adoption* of data science roles and methods in the short term.

[12] https://www.gse.harvard.edu/ppe/program/leadership-institute-academic-librarians
[13] https://www.gse.harvard.edu/ppe/program/library-leadership-digital-age

The broader impact of the National Forum workshop on the four factors highlighted above, will be measured by qualitative and quantitative methods during and after the workshop. Whilst innovation will be embedded and integral in the workshop design, the novel and pioneering applications of data science within libraries and institutions which are identified by participants, will be a major outcome. The exchange of ideas, experiences and opinions between the diverse participants, together with their projected dissemination within and beyond their home institutions, will facilitate wider knowledge transfer, ensuring that the data science community in libraries can grow, expand and consolidate in the future. An important component of the participant discussions will be data science skills and how they can be applied by librarians. For example, descriptions of requirements for real-world data science positions and iSchool curricula can be mapped against current and emergent library opportunities identified by administrators and managers, to assess educational relevance and value. The workshop will identify strategies for applying and integrating data science methods, workflows and tools in a range of library settings which will be distilled within the Roadmap. Participants' post-workshop activities will be investigated to measure the practical effectiveness of the workshop and to assess its value in fostering the strategic and operational adoption of data science methods and concepts.

Furthermore, the final report reflecting on the workshop will provide a resource for the broader library and data science communities. Given the limited scope of a one-year national forum grant, the workshop and the final report will be positioned as preliminary outputs designed to initiate conversations and to provoke further research and/or institutional commitment towards creating a culture of data science in libraries.



Incorporating data science into libraries will have a significant impact at multiple levels. Librarians who are technically literate and managers who can use data effectively will each act as catalysts to accelerate the acquisition of tangible benefits from data science when applied to library services, since the skills help them to perform their jobs better (Yelton 2015). By aligning with the open data and open science movements, libraries can break down functional barriers to allow the free and open movement of information and relationships throughout the organization. This creates a collaborative environment where people are making connections across the layers in the library, throughout the campus, and outwards into the local, national and global communities.

Within their campus communities, liaison libraries with technical skills and expertise in data-intensive research will better serve and collaborate with researchers across the disciplines. Librarians can help researchers with data sharing (Williams 2013), navigate increasingly complex data policy landscapes (Diekema et. al 2014), train graduate students and faculty with the skills to do data-intensive research.[14] Data science capability will also enable adoption of new immersive service delivery models where librarians collaborate on research projects as integral members of science teams (Lyon 2016).

Librarians are strategically positioned to work with their local communities around issues of data literacy. For example, librarians at the University of Pittsburgh have worked with the city and the county in the development of the Western Pennsylvania Regional Data Center, the region's civic open data platform, and leveraged their expertise in metadata, privacy, and data training.[15] The IMLS funded *Data Literacy for High School Librarians* project supports librarians to improve the pipeline of data literate citizens by engaging students in public schools.[16]

At the national and international level, we need technically skilled librarians to contribute to the development of our national digital platform. Librarians with these deep technical and data skills can control their own infrastructural destiny. Libraries have a rich history and knowledge about questions of ethics, privacy, and access, and as librarians and library administrators acquire and apply data science expertise, we want to ensure careful attention is paid to contexts within which this expertise is exercised. There is significant potential in leveraging data and advanced analytics in the management and operation of libraries, but we must also be careful not to relinquish agency to metrics. We also believe the broader data science community has much to learn from librarians' nuanced understanding of the social and technical dynamics around these issues.

3. Project Design

Our goal is to connect diverse, but disparate, communities working in and around libraries and data science. This specifically includes bringing together continuing educators to reflect on their experiences and develop a broader network. We bring educators in direct conversation with library managers and administrators to begin the conversation about what opportunities exist and what *structural changes* can be done to make data science training and application more effective. We bring industrial data scientists, whose experience in the application of data science to real world problems, will be a boon for librarians struggling to incorporate data science into their organizations and individual practice. We also aim to facilitate a conversation about how librarians can inform industrial data science about the growing social and community issues around data, especially around questions of ethics and access.

[14] https://github.com/DSSatPitt/katz-python-workshop

[15] https://www.cni.org/topics/digital-curation/libraries-will-be-an-asset-for-us-emerging-roles-for-academic-libraries-in-civic-data-partnerships

[16] http://datalit.sites.uofmhosting.net/

At the current time, the plethora of opportunities and promises around data science need to be *thoughtfully* and *strategically* applied in libraries. The objectives of this project are to:

- Better understand the workforce *skills* requirements and management gaps at both strategic and operational levels.
- Develop an 'Implementing Data Science Roadmap' with clear pathways for incorporating the *innovative* possibilities of data science in libraries.
- Identify operational contexts both internal and external to the library, where the *adoption* of data science methods may be most-effectively applied and have a measurable impact.
- Bring librarians into broader conversations in the data science community to *transfer* their awareness and knowledge of data privacy, ethics, and access into new domains.

To accomplish these goals, the project will organize a multi-day workshop to be held in the spring of 2017 at the University of Pittsburgh. This activity will be executed in three phases, a planning and preparation phase, the multi-day workshop, and an assessment, reporting and dissemination phase.

3.1 Planning & Preparation

Phase One has two goals: identification of participants and developing materials for the workshop. We will identify 25 to 30 experts from diverse domains related to data science and libraries. We will draw individuals from the following areas:

- *Library managers and administrators* - These are individuals responsible for leading libraries and teams, with an understanding of the organizational dynamics, writing and policy setting involved with data-intensive librarianship.
- *Formal library science educators* - This group includes faculty in LIS programs or iSchools who teach data-intensive courses and are currently training the next generation of data-savvy librarians.
- *Informal continuing educators* - This group includes members of the continuing education community, like Software Carpentry or our efforts in DST4L, focused on training the existing generation of librarians' data skills.
- *Practicing librarians -* This group focuses on practicing librarians who do data science and have experience working on and supporting data intensive research.
- *Data intensive faculty and researchers* - Individuals working on data-intensive research who have or could benefit from the support of and collaboration of librarians skilled in the methods of data science.
- *Industrial data scientists* - Practicing data scientists in industrial and commercial settings have experience with the integration of data-driven decision making in organizational environments.
- *Stakeholders and policymakers* - This would include institutional policy-makers such as the Vice-Provost for Research, but also national and international organizations like the Association of Research Libraries, the Coalition for Networked Information and CODATA.
- *Funders* - Organizations such as IMLS and the Sloan Foundation may be brought together to facilitate conversations about investments at the intersection of libraries and data science.

We have gathered letters of support (see supplementary materials) from a range of stakeholders in the library, research, publishing, and data science communities. The project team has a wide network and we have had informal conversations with key stakeholders representing each of the areas listed above.

3.2 The Workshop

The project organizers will convene a one-and-a-half day workshop hosted at the University of Pittsburgh. The workshop itself will be divided into two parts. The first half-day, will be an "Executive Briefing" session on data science for library management followed by a second, full-day, workshop on the opportunities and issues of data science in libraries. The briefing on the first day provides a conceptual grounding for the interactive discussions on the second day.

The half-day Executive Briefing will introduce participants to a set of case-studies focused on success stories on how data science has been effectively leveraged in the library. This briefing will feature two in-depth case-studies. The first will be focused inward exploring the impact of data science within the library, such as data-driven decision making, and managing a data driven organization. The second case study will be more outward facing, discussing the impact that librarians with data science skills can have outside the library in their communities. To craft these case studies, we will draw upon the project team's experiences as well as those of specific invited participants.

On the second day, invited experts will participate in structured conversation, brainstorming, and problem solving activities addressing the management and skills gap in libraries. The focus of this section of the workshop will revolve around the following questions:

1. What are the technical skills and roles needed to effectively leverage data science in libraries? Is there a single set of skills or an ecology of expertise?
2. How can the existing community of continuing education programs work in concert to more effectively share resources, expertise, and technical infrastructure? How tightly or loosely should they coordinate and what communities are currently being left out?
3. What are the data science opportunities for librarians both within and beyond the library? How can these opportunities be realised? What are the barriers and drivers to success?
4. What can continuing education programs do to address the management gap and support data-savvy librarians beyond individual workshops?
5. What is a compelling vision for data science librarianship that activates library administrators towards effectively and critically leveraging data science for decision making?

The project organizers will document the discussions and encourage collaborative note-taking, which, in addition to the materials used in the preparation of the workshop and the case-studies, will be used in the compilation of the final report discussed in the Communications Plan below.

Post-workshop activities for the remainder of the grant period will focus on writing a report and developing the "Implementing Data Science Roadmap." The primary audience for these documents will be members of the library community, especially managers and administrators. The report distills the conversations of the workshop and makes them available to a wider audience. The Roadmap lays the foundations of a strategic plan for data science in libraries. The foundation established by the report and Roadmap will used as part of larger capacity building efforts and be a resource for future applications to IMLS and other funding agencies.

After the forum we will post a survey to elicit feedback and to provide participants with additional opportunities to share information and evaluate the effectiveness of the workshop. We will provide workshop participants with opportunities to contribute to the final report and the Roadmap by providing case studies and success stories (through case-study templates for these contributions), about their experience incorporating knowledge from the workshop into their organizations.

4. Diversity Plan

As data science expands further into new sectors, the participation of women, underrepresented minorities, and persons with disabilities becomes ever more crucial. Currently, these populations are not sufficiently represented in the data science community. Programs such as Data Science for Social Good[17] and initiatives like DataKind[18] explicitly address issues of diversity, access, and representation in the field of data science and its applications. We will seek to invite members of these organizations to participate in the workshop and provide a perspective on data science for the social good. In addition, we will ensure members of underrepresented communities are invited to the workshop. We will recruit student assistants from under-represented populations to support and participate in the discussions.

As data science intersects with librarianship, questions of diversity, ethics, and access must be a priority in the conversation. In the workshop, beyond the skills and management gaps, we will be addressing issues of ethics, access, and diversity in and around data science. The field of data science has much to learn much from librarians' rich history and understanding of these issues. By bringing data scientists from industry into the conversation, we hope to facilitate dialogue about the impact of data on underrepresented communities and issues of diversity and access; the transformative potential of data science must not be seen to contribute to greater political, economic, and cultural divides.

5. Project Resources: Personnel, Time, Budget

5.1 Project Team and Contractors

Dr. Matt Burton is a Visiting Assistant Professor and Post-Doctoral Researcher at the School of Information Sciences & University Library System at the University of Pittsburgh. He has published on LIS education, Digital Humanities, and the use of data in Libraries. He has led multiple workshops training librarians, graduate students, and faculty from a variety of disciplines in data science methods. Dr. Matt Burton will dedicate a 5% effort towards all project activities and will act as the Project Manager.

Dr. Liz Lyon is a Visiting Professor at the School of Information Sciences and has worked in data curation and data science arenas for over a decade, previously as Associated Director of the UK Digital Curation Centre. She has published on emerging data science roles, and has substantive experience of facilitating consultation workshops addressing data-intensive science, data curation and university libraries. Dr. Liz Lyon will be dedicating a 5% effort towards all project activities.

Bonnie Tijerina is a researcher at Data & Society Research Institute in New York City working on grant-funded projects related to big data research ethics support on university campuses and online data privacy literacy in libraries. She has worked in libraries for over ten years and is founder and president of Electronic Resources & Libraries, a professional development organization. Bonnie Tijerina will be dedicating a 5% effort towards the workshop and writing the final report.

[17] https://dssg.uchicago.edu/

[18] http://www.datakind.org/

Chris Erdmann is the Head Librarian of the Harvard-Smithsonian Center for Astrophysics where he has developed the Data Scientist Training for Librarians based on his own work with the NASA Astrophysics Data System (ADS) and the astronomy community.

5.2 Budget

We estimate a total of $97,911 to convene a meeting at the University of Pittsburgh and to author a media-rich report of the workshop findings. Salaries and wages at 5% effort plus fringe benefits will be [redacted]. Invited participants will be geographically diverse, including some international partners. An allocation of [redacted] will be used to cover workshop participants' travel, room and board, and compensation for a maximum of 30 participants and [redacted] for conference and report writing for a total of [redacted] related travel expenses. A sum of [redacted] will allocated to the costs of consulting partners in authoring and publishing the report on the web. Student support and assistance at the workshop will be [redacted]. An additional [redacted] will cover indirect costs at the [redacted] negotiated rate.

5.3 Timeline

Phase	Timeframe	Activity
Planning	Fall / Winter 2016	● Develop attendee list ● Create workshop agenda ● Develop executive case-studies ● Report pre-work
Workshop	Spring 2016	● Run executive briefing ● Run workshop discussions
Assessment, Reporting & Dissemination	Summer / Fall 2016	● Follow-up with participants ● Compile notes ● Report Writing ● Present findings at relevant venues

Given the distributed nature of the team, the project manager will facilitate the project online through various collaboration tools (GitHub, Google Drive, Skype). We will coordinate through monthly and bi-weekly Skype meetings depending on the work phase. We have also specifically budgeted to bring the team to Pittsburgh for a multi-day writing sprint after the workshop.

6. Communications Plan

The reporting and dissemination phase will involve authoring a report and Roadmap summarizing and distilling the discussion for the broader community. This report will be published on the web for enriched media and broader community engagement (such as annotation with Hypothes.is) and deposited in the University of Pittsburgh's institutional repository, D-Scholarship.[19] The materials developed during the first phase, such as those for Data Science Training for Library Administrators, will be posted on GitHub

[19] http://d-scholarship.pitt.edu/

under an open access license. The final report and Roadmap will be distributed to workshop participants to inform potential transformational change.

Given the nature of the forum, the final report and outputs will need to include a data-centric analysis. The inward-facing case study will be built upon experience running a library and will draw upon data generated by library operations. Additional data collection and analysis performed during the workshop preparation and report writing phases of the project. The data, analysis, and visualization of these operational data will be published in the form of Jupyter Notebooks, a platform for composing and sharing data-rich computational narratives.

To ensure the report and findings reach the relevant constituencies, we will be presenting at relevant conferences such as the Coalition for Networked Information membership meetings, the Digital Library Federation forums, and the International Digital Curation Conference. Beyond library and academic venues, we will also present our findings to the data science community at venues such as the Open Data Science Conference[20] and the O'Reilly Strata conference.[21]

[20] https://www.odsc.com/

[21] http://conferences.oreilly.com/strata

Schedule of Completion

Task	November	January	March	May	July	September	November	Phase
Project Start (November 1st 2016)	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥 Preparation
Develop participant list	🟥							🟦 Workshop
Issue participant invitaitons	▨	🟥						🟨 Assessment, Reporting & Dissemination
Finalize participants		🟥	🟥					
Develop workshop materials	▨	🟥	🟥	▨				
Finalize workshop Schedule			🟦					
Workshop (dates TBD, Sprint 2017)			▨	🟦	▨			
Issue feedback survey				🟦	🟦			
Draft intial report			▨	🟨				
Circulate draft report w/ participants				🟨	▨			
Incorporate participant feedback					▨	🟨		
Publish report online						🟨	▨	
Disseminate final report								

Bibliography

Burton, M. (2016). Looking for the Core: Preliminary Explorations of iCaucus Syllabi. Presented at the IConference 2016 Proceedings, iSchools. http://doi.org/10.9776/16225

Diekema, Anne R., Andrew Wesolek, and Cheryl D. Walters. (2014) “The NSF/NIH Effect: Surveying the Effect of Data Management Requirements on Faculty, Sponsored Programs, and Institutional Repositories.” The Journal of Academic Librarianship 40 (3-4): 322–31. doi:10.1016/j.acalib.2014.04.010.

Lyon, L. (2016) Librarians in the Lab: Toward Radically Re-engineering Data Curation Services at the Research Coalface. New Review Academic Librarianship. http://www.tandfonline.com/doi/abs/10.1080/13614533.2016.1159969

Lyon, L., & Mattern, E. (2016) Education for Real-World Data Science Roles (Part 2): A Translational Approach to Curriculum Development. IJDC (In Press).

Lyon, L., & Brenner, A. (2015) Bridging the Data Talent Gap: Positioning the iSchool as an Agent for Change. International Journal of Digital Curation, 10(1), 111–122. http://doi.org/10.2218/ijdc.v10i1.349

Mayernik, Matthew S., Lynne Davis, Karon Kelly, Bob Dattore, Gary Strand, Steven J. Worley, and Mary Marlino. (2014) “Research Center Insights into Data Curation Education and Curriculum.” In Theory and Practice of Digital Libraries – TPDL 2013 Selected Workshops, edited by Łukasz Bolikowski, Vittore Casarosa, Paula Goodale, Nikos Houssos, Paolo Manghi, and Jochen Schirrwagen, 416:239–48. Communications in Computer and Information Science. Cham: Springer International Publishing. doi:10.1007/978-3-319-08425-1.

Williams, Sarah C. 2013a. “Data Sharing Interviews with Crop Sciences Faculty: Why They Share Data and How the Library Can Help.” Issues in Science and Technology Librarianship 72. doi:10.5062/F4T151M8.

Yelton, A. (2015). Coding for Librarians: Learning by Example. Library Technology Reports, 51(3). http://doi.org/http://dx.doi.org/10.5860/ltr.51n3

Council for Big Data, Ethics, and Society. (2016) “Perspectives on Big Data, Ethics, and Society.” Accessed May 27, http://bdes.datasociety.net/council-output/perspectives-on-big-data-ethics-and-society/.

DIGITAL STEWARDSHIP SUPPLEMENTARY INFORMATION FORM

Introduction

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded research, data, software, and other digital products. The assets you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. However, applying these principles to the development and management of digital products is not always straightforward. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and best practices that could become quickly outdated. Instead, we ask that you answer a series of questions that address specific aspects of creating and managing digital assets. Your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions

If you propose to create any type of digital product as part of your project, complete this form. We define digital products very broadly. If you are developing anything through the use of information technology (e.g., digital collections, web resources, metadata, software, or data), you should complete this form.

Please indicate which of the following digital products you will create or collect during your project (Check all that apply):

	Every proposal creating a digital product should complete …	Part I
	If your project will create or collect …	**Then you should complete …**
☑	Digital content	Part II
□	Software (systems, tools, apps, etc.)	Part III
□	Dataset	Part IV

PART I.

A. Intellectual Property Rights and Permissions

We expect applicants to make federally funded work products widely available and usable through strategies such as publishing in open-access journals, depositing works in institutional or discipline-based repositories, and using non-restrictive licenses such as a Creative Commons license.

A.1 What will be the intellectual property status of the content, software, or datasets you intend to create? Who will hold the copyright? Will you assign a Creative Commons license (http://us.creativecommons.org) to the content? If so, which license will it be? If it is software, what open source license will you use (e.g., BSD, GNU, MIT)? Explain and justify your licensing selections.

In accordance with University of Pittsburgh policy 11-02-02, the copyright of all scholarly works resides with the authors. We will assign a Creative Commons By Attribution 4.0 International (CC-BY) http://creativecommons.org/licenses/by/4.0/

A.2 What ownership rights will your organization assert over the new digital content, software, or datasets and what conditions will you impose on access and use? Explain any terms of access and conditions of use, why they are justifiable, and how you will notify potential users about relevant terms or conditions.

We derive conditions of use from the CC-BY license, where derivative works must include an attribution, appropriate credit, to the original work. Long-term, public, and open access to the report will be maintained through D-Scholarship@Pitt, the University of Pittsburgh's institutional repository.

A.3 Will you create any content or products which may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities? If so, please describe the issues and how you plan to address them.

A draft of the final report that is derived from the workshop will be circulated with participants before being published.

Part II: Projects Creating or Collecting Digital Content

A. Creating New Digital Content

A.1 Describe the digital content you will create and/or collect, the quantities of each type, and format you will use.

The report will be published as a dynamic website, GitHub repository, and static PDF document.

A.2 List the equipment, software, and supplies that you will use to create the content or the name of the service provider who will perform the work.

Content will be hosted on Github.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to create, along with the relevant information on the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

The site will be published in responsive HTML, CSS, and Javascript and will include standard formats for web-images.

B. Digital Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance (e.g., storage systems, shared repositories, technical documentation, migration planning, commitment of organizational funding for these purposes). Please note: You may charge the Federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the Federal award. (See 2 CFR 200.461).

C. Metadata

C.1 Describe how you will produce metadata (e.g., technical, descriptive, administrative, or preservation). Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, or PREMIS) and metadata content (e.g., thesauri).

C.2 Explain your strategy for preserving and maintaining metadata created and/or collected during and after the award period of performance.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of digital content created during your project (e.g., an API (Application Programming Interface), contributions to the Digital Public Library of America (DPLA) or other digital platform, or other support to allow batch queries and retrieval of metadata).

D. Access and Use

D.1 Describe how you will make the digital content available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

D.2 Provide the name and URL(s) (Uniform Resource Locator) for any examples of previous digital collections or content your organization has created.

Part III. Projects Creating Software (systems, tools, apps, etc.)

A. General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) this software will serve.

A.2 List other existing software that wholly or partially perform the same functions, and explain how the tool or system you will create is different.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your software (systems, tools, apps, etc.) and explain why you chose them.

B.2 Describe how the intended software will extend or interoperate with other existing software.

B.3 Describe any underlying additional software or system dependencies necessary to run the new software you will create.

B.4 Describe the processes you will use for development documentation and for maintaining and updating technical documentation for users of the software.

B.5 Provide the name and URL(s) for examples of any previous software tools or systems your organization has created.

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under an open-source license to maximize access and promote reuse. What ownership rights will your organization assert over the software created, and what conditions will you impose on the access and use of this product? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain any prohibitive terms or conditions of use or access, explain why these terms or conditions are justifiable, and explain how you will notify potential users of the software or system.

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

C.3 Identify where you will be publicly depositing source code for the software developed:

Name of publicly accessible source code repository:
URL:

Part IV. Projects Creating a Dataset

1. Summarize the intended purpose of this data, the type of data to be collected or generated, the method for collection or generation, the approximate dates or frequency when the data will be generated or collected, and the intended use of the data collected.

2. Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

3. Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

4. If you will collect additional documentation such as consent agreements along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

5. What will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

6. What documentation (e.g., data documentation, codebooks, etc.) will you capture or create along with the dataset(s)? Where will the documentation be stored, and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

7. What is the plan for archiving, managing, and disseminating data after the completion of the award-funded project?

8. Identify where you will be publicly depositing dataset(s):

Name of repository:
URL:

9. When and how frequently will you review this data management plan? How will the implementation be monitored?

Original Preliminary Proposal

The Data Scientist as the 21st Century Librarian?

Data science is emerging as a powerful mechanism for "generating insight from data to inform decision making"[1] and is transforming organizations in all sectors from commerce, to healthcare, to entertainment. As librarians increasingly use, build, and maintain the National Digital Platform, the skills to manipulate, analyze, and manage data are crucial. Librarians will need to leverage the tools and techniques of data science, but they currently face two significant challenges:

1. *The Skills Gap* - While practicing mid-career librarians are learning some data science skills, it is through ad-hoc, uncoordinated continuing education programs.
2. *The Management Gap* - Library administrators need toolkits and frameworks to strategically leverage data science for data-driven decision making and management of library operations.

These mutually constitutive gaps create disconnects in the application of data science for librarianship. To address this problem the University of Pittsburgh School of Information Sciences and the Harvard/Smithsonian Center for Astrophysics request $83,373 to convene experts from inside and outside the library community to articulate a vision and roadmap for data science in libraries.

Objectives & Motivations

Our goal is to coalesce diverse and disparate communities whose experiences and perspectives on data science inspire and contribute to *developing* and *sustaining* our National Digital Platform. We are at a critical moment where a plethora of opportunities and promises around data science need to be *thoughtfully* and *strategically* applied in libraries. Our objectives are to:

- Articulate a vision for data science in libraries balancing data's transformative potential with a critical awareness of privacy, equality of access, and ethics in data-rich environments.
- Develop a roadmap with clear pathways for developing the next generation of continuing education programs focused on data science skills for librarians.
- Provide guidance and recommendations for preparing the next generation of library administrators on managing data intensive teams and organizations.

This proposal builds upon experiences running Data Science Training for Librarians(DST4L) a program started in 2012 at the Harvard/Smithsonian Center for Astrophysics. DST4L is a multi-day workshop for academic librarians teaching data analysis, visualization, and the research lifecycle. DST4L is one of many similar programs whose aim is to address the technical *skills gap* of mid-career librarians.[2]

A lack of coordination across these programs has resulted in an uneven landscape of orientation and objectives. DST4L focuses on data science skills (data hacking, analysis, visualization); Software Carpentry emphasize software development skills (programming, engineering) for building tools and infrastructure; Library Carpentry, focuses on data wrangling (cleaning and formatting) and leaves analysis to the domain experts. This community needs an opportunity to come together to share their experiences and develop a collaborative vision.

Our experiences with DST4L has shown while attendees express enthusiasm in learning technical skills, when they return to their home institutions they cannot effectively apply their newly acquired skills in the course of their work. Without the appropriate organizational and technical infrastructures in place, a practicing librarian's expertise cannot reach across *the management gap.* Conversely, Library administrators are not benefitting from data science capability and libraries are not realising the full potential and power of data science through the cultivation of new work practices.

Our goal is to address the skills and management gaps by bringing together diverse constituencies to discuss opportunities for coordination and collaboration. Overcoming these gaps is

[1] http://christianlauersen.net/2016/01/11/librarians-as-data-scientists/

[2] http://altbibl.io/dst4l/, http://software-carpentry.org/, http://www.datacarpentry.org/, http://librarycarpentry.github.io/, http://andromedayelton.com/

crucial for connecting and interconnecting the National Digital Platform to the other domains already transformed by data science.

National Forum Grant Activities

This grant will be executed in three phases. In Phase 1, the investigators will identify 25 to 30 experts from diverse domains related to both data science and libraries who we recognize as having related goals and complementary expertise. This group includes practitioners and administrators in libraries, continuing educators (such as Software and Library Carpentry), faculty and researchers of data science and data studies (such as Data and Society Research Institute), stakeholders in national organizations (such as OCLC and DLF), funders of data science initiatives (such as the Sloan and Moore Foundations), and data scientists from industry (such as O'Reilly Media and GitHub).

In the second phase we will convene a multi-day consultation workshop at the University of Pittsburgh. The focus of the workshop will revolve around the following questions:

1. What are the technical skills and roles needed to effectively leverage data science in libraries? Is there a single set of skills or an ecology of expertise?
2. How can the existing community continuing education programs work in concert to more effectively share resources, expertise, and technical infrastructure? How tightly or loosely should they coordinate and what communities are currently being left out?
3. What can these programs do to address the management gap and support data savvy librarians beyond individual workshops?
4. What is a compelling vision for data science librarianship that activates library administrators towards effectively and critically leveraging data science for decision making?

These questions will be explored through breakout sessions, brainstorming activities, expert lectures, and structured exercises. Organizers will document the discussion and encourage collaborative notetaking.

The third and final phase will involve authoring a report summarizing and distilling the discussion for the broader community. This report will be published on the web for enriched media and broader community engagement (i.e. annotation with Hypothes.is). The report will not only be a document, but a focal point for a continued conversation around data science and libraries.

Budget

We estimate a need of $83,373 in order to convene a meeting at the University of Pittsburgh and author a media rich report. Invited participants will be geographically diverse, including some international partners. An allocation of around [redacted] will be used to cover travel, room and board, and compensation for a maximum of 30 participants; [redacted] is allocated for consulting costs for authoring and publishing the report on the web; [redacted] will be used to organizer's cover travel, time, and student support. [redacted] would cover indirect costs at [redacted].

Project Team

Dr. Liz Lyon is a Visiting Professor at the School of Information Sciences and has worked in data curation and data science arenas for over a decade, previously as Associated Director of the UK Digital Curation Centre. She has published on emerging data science roles, and has substantive experience of facilitating consultation workshops addressing data-intensive science, data curation and university libraries.

Dr. Matt Burton is a Visiting Assistant Professor and Post-Doctoral Researcher at the School of Information Sciences & University Library System at the University of Pittsburgh. He has published on LIS education, Digital Humanities, and the use of data in Libraries.

Chris Erdmann is the Head Librarian of the Harvard-Smithsonian Center for Astrophysics where he has developed the Data Scientist Training for Librarians based on his own work with the NASA Astrophysics Data System (ADS) and the astronomy community.