LIS Education and Data Science-for-the National Digital Platform (LEADS-4-NDP)

Abstract

As purveyors and stewards of information, library scientists and archivists have important and growing responsibilities to actively and meaningfully participate in our data-driven world, and apply data science techniques to improve library services and operations. It is essential, thus, to educate the next generation of Library and Information Science (LIS) faculty in tools, skills, and service needed for the National Digital Platform (NLP) and other data-driven services. Despite this need, LIS data-focused graduate programs primarily address digital curation, with limited attention to analytical skills. This is not surprising, given that only 3% (11 of 330) schools on Swanson's Awesome Data Science Colleges list (2016), include information science partnered with library science. In response to these challenges, we propose this project, **LIS Education and Data Science for the National Digital Platform (LEADS-4-NDP)**, as part of the Laura Bush 21st Century Librarian Program, to address pressing challenges in preparing the next generation of LIS faculty who will need to integrate data science and LIS education.

The **LEADS-4-NDP** program presents a unique educative program that integrates library science with recent data science advancements. The program will provide stipends for **18 LIS doctoral students**, as **LEADS Fellows**, from ALA accredited programs across the country. LEADS Fellows, will complete: **1)** an online preparatory curriculum, **2)** an intensive 3-day data science bootcamp at Drexel University, and **3)** a ten-week data science internship with a LEADS-4-NDP project partner. This program leverages the data science doctoral program curriculum at the College of Computing and Informatics, Drexel University (Drexel, 2016). LEADS-4-NDP will build a national cohort among the next generation of LIS faculty and bring necessary knowledge and skills in data science to LIS education. A diversity plan includes working with the ALA's Office for Diversity, Literacy, and Outreach Service to recruit broadly in LIS, the Drexel Dornsife Center which convenes programming for a diverse community by weaving together Drexel's knowledge resources with the expertise of community-based partners, and the NDP partnership—Mapping Inequality project, showing how data science can be applied to study race and diversity in NDP collection materials.

Key goals of the LEADS-4-NDP are to:

1. Enhance current LIS doctoral curricula through targeted data science education and an immersive research experience with leading NDP partners.
2. Form a national cohort among LEADS participants and future LIS faculty, who will be able to bring new data science knowledge and skills to LIS education on a national scale.
3. Build educational infrastructure, including course materials, lesson plans and shared datasets that LEADS-4-NDP participants and LIS faculty will be able to access and use for research and LIS graduate education.
4. Develop a sustainable educative model encompassing diversity, and which can be expanded to include additional NDP partners and LIS programs.

LEADS-4-NDP aligns with IMLS/LB21 goals and the NDP's need for a digital library workforce having an expanded skill-set (IMLS, 2015). It will advance doctoral student learning capacity, knowledge, and skill on a national scale via cross-training, hands-on experience, and collaboration. The project will have a national impact in four key areas: 1) LIS doctoral education enhancement, 2) Nationwide doctoral student participation, 3) NDP partner engagement from coast-to-coast, and 4) Developing a vetted, cost-effective learning model of data science for LIS students.

Drexel University submits this proposal to the Laura Bush 21st Century Librarian Program (LB21)—Doctoral-level program category, with a focus on the National Digital Platform (NDP). In collaboration with leading NDP organizations (identified in Section 2, Project Design), we propose the **LIS Education And Data Science for the National Digital Platform** (LEADS-4-NDP) program, a unique summer experience to enhance LIS doctoral curricula in the area of data science. The program will support **18 LEADS Fellows** from across the country, providing immersive internship experiences with NDP partners. LEADS-4-NDP takes important steps toward building a national cohort among the next generation of LIS faculty, imparting necessary data science knowledge and skills to LIS education.

1. STATEMENT OF NEED

As purveyors and stewards of information, library scientists and archivists have important and growing responsibilities to actively and meaningfully participate in our data-driven world, and apply data science techniques to improve library services and operations. Data science leverages data analytic techniques to gain new insights, model and test approaches, and improve overall system performance (Showers, 2015; Stanton, 2012). To this end, it is essential to educate the next generation of library and information science (LIS) faculty in data science methods, skills, and tools (Dumbill, et al, 2013; Stanton, 2012). Education in this area is vital to optimize and sustain the growing "digital capability and capacity of libraries and museums across the US" (IMLS, 2015) encompassing the National Digital Platform (NDP). Despite this pressing need, LIS data-focused graduate education today primarily focuses on digital curation, with limited attention to analytical skills as part of the core competencies. This situation is not surprising, given that only 3% (11 of 330) schools on Swanson's Awesome Data Science Colleges list (2016), include programs partnering with library and information science.
In this project, we seek to address two of the most pressing challenges evident in library and information science (LIS) impacting the ability to leverage data science. These challenges are:

1. **A shortage of LIS faculty prepared to teach data science in the LIS context.**
2. **An absence of educational infrastructure for teaching data science in LIS programs.**

These challenges are reflected in current LIS curricula and underscore the following needs:

Need for faculty preparation. LIS doctoral students, specifically those who seek faculty positions in ALA accredited programs, need practical and methodological exposure to data science in order to advance LIS curricula. To date, data training for future LIS faculty has primarily focused on digital curation and preservation (Heidorn, 2011), and data management in disciplines (Tenopir, et al., 2016) such as earth science, biology, health, and social science (Bieraugel, 2016). LIS faculty need to learn how to formulate data science questions that will aid library assessment and daily operations; they need to learn how to select the appropriate data science methodologies, evaluate the results, and present the results to stakeholders. Through such training they will be able to better teach and prepare the next generation of librarians who will grapple with the growing big data sources underlying libraries and other information centers. LIS faculty need to be acquainted with the trove of data science tools and resources, including data cleansing software and tools, descriptive and predictive data analytics techniques, machine learning methods, and data visualization software packages, among others (EMC Education Services, 2015). Finally, LIS doctoral students seeking faculty positions need immersive internships to gain first-hand experience with NDP big data. Immersive experiences will empower future faculty with data-driven problem-solving skills and knowledge to cope with big data challenges in LIS areas (Song and Zhu, 2016).

Need for LIS/data science educational infrastructure. The field of LIS needs to build an educational infrastructure around the National Digital Platform (NDP). LIS has a robust, educational infrastructure with student access to systems such as ProQuest for learning 'search,' and OCLC's Connexion and the Library of

Congress' Cataloger's Desktop for learning cataloging. Additionally, these areas and other LIS foci have textbooks, online tutorials, and other resources forming a framework for education. Initial infrastructure steps have been taken primarily in continuing education. Examples include the Data Scientist Training for Librarians (DST4L, 2014), developed for libraries at the Harvard-Smithsonian Center for Astrophysics, John G. Wolbach Library and the Harvard Library, which presents an exemplary LIS/data science curriculum (Erdmann, 2015). IMLS also recognizes this need through recent support for a national forum at the University of Pittsburgh School of Information Sciences, where a group of experts from within and outside the library community gathered to identify skills and management gaps for integrating data science into libraries, and to inform continuing education (IMLS, 2016). Supporting this is the work of Lyon, L., & Brenner (2015), Lyon, L., & Mattern (2016), recognizing current skills gaps, and leading work explaining the growing digital landscape Borgman, 2015. Additionally, the work of Drexel University has also taken important steps in this area with the 2016 revision of its information studies PhD program, which now includes data science as one of three LIS doctoral curricula specializations. Drexel's leadership in this area, and the other cases noted, underscore the necessity of a national, shared infrastructure comprising tools, techniques, and demonstrations supporting education and training in the LIS/data science area.

Need for LIS/data science community development. Interconnected with infrastructure, the field of LIS needs to seed LIS/data science community development. Similar to data curation efforts, where LIS graduate programs and future faculty have been supported through IMLS funded projects to learn and share ideas and develop a sustainable community, there is a need to invest in similar underpinnings to build a cohesive data analytics community for the LIS field. Without a community, LIS graduates who gain data science experience are at risk of leaving the discipline for industry (Google or Facebook, for example). Research shows that a connection to a community and a sense of place are reasons people chose to join and remain in a discipline (Serrat, 2017). Gaëta, et al (2017) outlines these as rules for scientific professional societies and community growth. Although seeking patterns and predicting performance is a common theme across data science components of many endeavors, LIS/data science focuses currently on services and operations, distinguishing the field from business, computer science, and other areas. Building a community that seeks to address these factors is essential to the success of data science application. Data science education within LIS programs must address fundamental data and information organization issues as well as using data to make decisions to improve libraries, archives, and information centers' day-to-day operations and service (Showers, 2015; Song and Zhu, 2016). Thus, it is critical to invest in community development by educating the next generation of LIS faculty members, so they can advance curriculum changes for students entering the field.

The challenges reviewed above motivate the proposed **LIS Education and Data Science for the National Digital Platform (LEADS-4-NDP)** program, informing the goals and deliberate steps that comprise our project design.

2. PROJECT DESIGN

The **LEADS-4-NDP** program presents a unique educative program that integrates library science with recent data science advancements. The program will provide stipends for **18 LIS doctoral students** from ALA accredited programs across the country. Participants, as LEADS Fellows, will complete: **1)** an online preparatory curriculum, **2)** an intensive 3-day data science bootcamp at Drexel University, and **3)** a ten-week data science internship with a LEADS-4-NDP project partner. This program leverages the data science doctoral program curriculum at the College of Computing and Informatics, Drexel University (Drexel, 2016). Drexel's curriculum provides graduates the ability to conduct LIS focused data science research, and those who pursue academia bring data science knowledge and techniques to LIS curricula.

2.1 Goals

Key goals of the **LEADS-4-NDP** program are to:

1. **Enhance current LIS doctoral curricula** through targeted data science education and an immersive research experience with leading NDP partners.
2. **Form a national cohort among LEADS participants and future LIS faculty**, who will be able to bring new data science knowledge and skills to LIS education on a national scale.
3. **Build educational infrastructure**, including course materials, lesson plans and shared datasets that LEADS-4-NDP participants and LIS faculty will be able to access and apply to research and LIS graduate education.
4. **Develop a sustainable educative model** encompassing diversity, and which can be expanded to include additional NDP partners and LIS programs.

LEADS-4-NDP is modeled, in part, on the DataONE datanet summer internship program and the Research Data Alliance. The proposed program specifically draws from the successful model supporting virtual internships; although LEADS-4-NDP is unique in its focus on LIS doctoral students as future educators and the NDP. LEADS-4-NDP PI, Professor Jane Greenberg, has been substantively engaged in both of these internship programs as an instructor and mentor, and brings important expertise to the proposed program. The sub-sections that follow provide additional project design details covering plans and activities, personnel and management, communication and outreach, evaluation and outcome assessment, and risk mediation.

2.2 Plans and Activities

LEADS-4-NPD program **plans and activities** cover four phases, a set of learning objectives, and curriculum modules so that the NDP internships meet program goals.

LEADS-4-NPD Project phases

- **Phase 1 – Curriculum & internship development (June 2017-December 2017)**

Project PIs will develop curriculum modules covering 1) Introduction to data science, 2) Digital curation and metadata for data science, 3) Analytical algorithms and data analytics systems, 4) Data Visualization, 5) NDP projects and services, and 6) Developing data science LIS lesson plans. One component will be online, with a 3-day bootcamp held prior to the 10-week internship experience. During the development phase, PIs will collaborate with NDP partners to further refine NDP projects and link internship needs to the LEADS-4-NDP curriculum. Advisory board members will provide feedback on plans and assess project progress.

- **Phase 2 – First summer camp, preparation and execution (Summer 2018)**

Early in 2018, project PIs will prepare recruiting materials, engage LIS doctoral programs across the country to recruit students, connect with NDP partners and advisory board members to review camp applicants and match applicants to the appropriate internship project. Students will apply in late March 2018, and be notified of acceptance by mid-April 2018. Successful applicants, as **LEADS fellows**, will have access to the online curriculum in May, and will visit Drexel University in late May for the on-site data science bootcamp.

Mentors and advisory board members will join selected segments of the bootcamp so that mentors and mentees will have the opportunity to meet in person. Following this phase, students will continue via Blackboard, pursuing their NDP-data internships. LEADS fellows will also visit their NDP project sites during their internship experience. During Fall 2018, LEADS fellows, mentors, and advisory board members will gather at a selected conference (e.g., the Coalition for Networked Information (CNI) fall conference) to hold an open

forum and share project outcomes. PIs will also use this time to evaluate the first summer camp, drawing from the IMLS evaluation framework. Our approach will include surveys, focus groups and interviews with LEADs fellows, advisory board members, and NDP project mentors.

- **Phase 3 – Second iteration and improvement (January 2019-August 2019)**

PIs will incorporate feedback and revise the LEADS-4-NDP curriculum for a second camp. PIs will work with NDP partners and seek feedback from advisory board members as they develop and refine year two (2019) data science internship experiences. LEADS Fellows will be recruited for the second summer camp, following a schedule similar to that used for the first cohort. During this phase, the first group of LEADS Fellows will be encouraged to disseminate results and learning experiences via ALA, ALISE, ASIST, and other relevant venues. The project budget allocates support for LEADS Fellows who seek to share their work at professional and scholarly conferences, and support for project mentors to join them in this experience.

- **Phase 4 –A sustainable LEADS-4-NDP model (September 2018-September 2019)**

At the end of the first summer camp, PIs will summarize lessons learned and accumulated knowledge will inform the development of a sustainability model that supports a national LEADS-4-NDP program. A national, sustainable program will include: **1)** yearly calls-for-participation (CfP) for NDP partners, along with project mentors, and **2)** broader engagement from LIS programs across the nation. Section 4, National Impact, provides background information on developing a sustainable framework. One long-term goal is to develop an educational model that is extensible and transferable to other LIS programs. The LEADS-4-NDP will revise, finalize, and share the model toward the end of the program. The model will be shared via an open forum planned for the fall 2019 CNI conference. All LEADS Fellows, and LEAD-4-NDP partner-mentors, advisory board members, and core project staff (PI and administrative assistant) will share project outcomes and experience at relevant LIS conferences. Core project staff will also disseminate project outcomes and assessment results via the scientific and other scholarly literature. Equally important, the LEADS team will make the model, curriculum, and lesson plans accessible online to the LIS program community both nationally and internationally through the Drexel University Metadata Research Center's website, providing easy access, while preserving materials in the IDEAS repository at Drexel University.

LEADS-4-NDP Curriculum: Learning Outcomes, Structure, and Internships

The LEADS curriculum targets fundamental aspect of data science, and will allow LEADS Fellows to develop necessary competency to pursue their internship experience. The curriculum has a flexible design, recognizing that LEADS Fellows will very likely have a diverse set of skills and background knowledge. The LEADS team anticipates that some LEADS Fellows will have little, if any programming background, while others may have gained a degree of competency with R or Python. Python will be the chief language for the program. The broad skillset of the four PIs will allow for adaptations as necessary.

The curriculum has six learning outcomes. Upon successful completion of the LEADS-4-DATA, LEADS Fellows will be able to:

1. Understand a number of major data science techniques.
2. Work with a set of available data science technologies and tools, and communicate their applicability to the NDP.
3. Demonstrate experience applying mathematics, statistics, machine learning, and data mining for preprocessing data, modelling, and invoking data science methods for problem solving.

4. Conduct data science experiments to discover knowledge and evaluate outcomes (for example, trend identification, visualization, and analytics).
5. Effectively communicate an LIS/data science scenario.
6. Confidently collaborate with NDP partners and domain experts in order to examine the underlying structure of data and to report with accuracy and insight.

The LEADS-4-NDP curriculum covers: 1) library data life cycle management, 2) data processing, analytics, and decision support, including data cleansing, transformation, mining, statistics, machine learning, predictive analytics, and visual analytics, and 3) data integration to support library systems and services. The following curriculum modules are based on the categories of essential data science skills and relevant processes in LIS. The curriculum is designed for exposure to a variety of fundamental topics with detailed attention to subjects necessary for the immersive internship. A supplemental document contains a more detailed version of this plan, including required curriculum reading.

Pre-requisites

- Knowledge of research methods and basic statistics (e.g., graduate level research methods and statistics coursework, or demonstration of same through current or previous project work).
- Experience with at least one of the following statistical packages: Excel, SPSSS, R, MATLAB, or SAS. Technical and programming skill and experience with at least HTML, Java script, Python, R, or Java. (As noted above, Python will be the primary camp language, and the pre-camp module will prepare students as necessary).

Table1: Online Learning Modules and Preparation

Online modules before the camp
Python for Data Science Introduces the use of Python and related packages for data science. Covers specific packages including NumPy, SciPy, and Scikit-learn.
Software Setup and Exercises Students learn and follow the tutorials for initial software setup and practice basic Python skills.
Preparation with Project Data and Domain Students communicate with mentors on domain and potential data, obtain data (samples), and understand data in the context of project domain.

Table 2: Three-day Training

Day 1: Data
Introduction to Data Science Lecture/discussions: Five Vs of big data and their implications, data-driven paradigms, data economy, basic data science techniques; challenges and opportunities in the LIS context. Mentor/LEADS Fellows introduce data science projects.
LIS Data and Data Management Lectures/discussions: Library data and information (e.g., collection, patrons and transactions), databases and systems for library services, structured and unstructured data, NoSQL data management.
Metadata, Data Quality, and Integration Lectures/discussions: Categories and usability of metadata, importance of data quality, and techniques for data integration. Activities: Data cleaning, metadata interoperability and parsing.

Day 2: Methods
Data pre-processing Activities: More techniques used in data, including cleansing and transformation.
Data Mining and Machine Learning I: Methods Methods for association, clustering, classification, and numeric predictions.
Data Visualization and Visual Analytics Lectures/discussions and activities: Methods and tools for data/information visualization, visual analytics and discuss interpretation and reporting of visualization, visualization-supported data exploration, mining, and decision making.
Day 3: Computing
Large-scale and Parallel Computing Lectures/discussions: Implications of big data (e.g., volume and velocity) and computing frameworks for scaling up and scaling out (e.g., with MapReduce/Hadoop/Spark).
Data Mining and Machine Learning II: Computing Lectures/discussions: Computing aspects of machine learning algorithms for data mining and issues such as efficiency, time/space complexity, and scalability. Activities: Machine learning algorithms for data mining
Cloud-based Automated Data Analysis Lectures/discussions: Cognitive computing and IBM Watson Analytics for cloud-based automated data analysis and related cloud computing resources. Activities: Practice with Watson Analytics for data visualization, data prediction, and report generation.
Data Clinics and Method Consultation Activities: In concluding the camp, we will conduct consultations with students on their data projects. The focus will be on problem solving with methods selection and computing tools. Discussion: Closing discussion on continuing communications.

Internships

A signature component of the LEADS-4-NDP Project Design includes immersive internships with NDP partners. Each student will address questions and work with a mentor and a faculty member. Table 2 lists NDP partner internship projects. LEADS Fellows will be matched to these projects according to their skill and experience. The application process (see supplemental document), will require students to identify and rank three desired internships, following an approach used both in the DataONE and RDA internship programs. Project mentors will engage in reviewing all applicants to determine the best fit for each internship.

Table 3: List of LIS Data Analytics Projects Proposed by Partners for the Project Participants

Partner Organization	Project title	Project Outcome
OCLC	Visualizing Library and Information Science Monographs	Knowledge of the current state of LIS literature, and potentially, networks of authorship.
OCLC	Library Collection and Circulation Analysis	Better knowledge of collection use patterns in academic libraries.
California Digital Library (CDL)	Building and Maintaining Metadata Vocabulary Terms	Increasing adoption among groups concerned with digital preservation, citizen science, and design safety.
Free Library of	Provenance of Provenance	Analytics protocol to track collection life-

Philadelphia		cycle use and popularity.
Free Library of Philadelphia	Visualizing Temporal Use, Predicting Service Needs	Visualization of temporal use to aid in system service needs.
Historical Society of Pennsylvania	Enhancing access to historic biographical data through visualization tools	Improving access to historic data for scholars and other user groups and the ability to visualize connections between people, places, and institutions.
DCIC, University of Maryland	Mapping Inequality	Contributing to the in progress and recently released Mapping Inequality platform (http://mappinginequality.us).
BHL, Smithsonian Libraries	Mining Georeferences from Biodiversity Literature	Responding to Geo-questioning for our planet's biodiversity through coordinating data from BHL and GBIF.
University of Pennsylvania	Extracting and normalizing medieval manuscript data	Improving access to OPenn visualization: Predicting the possible

NDP partner mentors will contribute to the curriculum design, with real cases, and mentor the students during the 10-week internship period. Project mentors represent leading NDP organizations; the mentors are listed below, and letters of commitment have been included as a supporting document attachments.

Confirmed mentors:

- Laurie Allen, Assistant Director for Digital Scholarship, University of Pennsylvania (U. Penn) Libraries
- Emily B. Gore, Director of Content, Curation & Use, Digital Public Library of America
- John Houser, Historical Society of Pennsylvania (HSP)
- Martin R. Kalfatovic, Program Director, Biodiversity Heritage Library, and Associate Director, Smithsonian Libraries, Smithsonian Institution
- John Kunze, Identifier Systems Architect, California Digital Library, University of Berkeley
- Michael J. Kurtz, Associate Director, DCIC, College of Information Studies, University of Maryland.
- John C Meier, Deputy Director, Digital Strategies and Information Technology, Free Library of Philadelphia
- Dot Porter, Curator, Digital Research Services, U. Penn Libraries
  Roy Tennant, Senior Program Officer, OCLC

2.3 Personnel and Management

Project personnel. Project staff include four faculty members in the Information Science Department, College of Computing and Informatics, Drexel University, and an Administrative Assistant. Drexel faculty members include: Jane Greenberg, Alice B. Kroeger Professor, and Director of the Metadata Research Center (MRC); Xia Lin, Professor, and Director of International Programs; Il-Yeol Song, Professor, and Weima Ke, Associate Professor. All four PIs teach in data science, and will serve as LEADS-4-NDP curriculum instructors. Greenberg has expertise in metadata, semantics, and data management; Lin has expertise in visualization, digital libraries, and data science; Song has expertise in database design and analytics; and Ke has expertise in information retrieval, algorithms, and data analytics. The Administrative Assistant will work closely with the project PIs to coordinate the doctoral student application process, help arrange doctoral student travel and accommodations for summer bootcamp, and assist in travel arrangements and virtual connectivity for advisory board members and mentors to attend bootcamp. (Faculty resumes and a job description for the Administrative Assistant are in supporting documents).

Management Lead PIs Greenberg and Lin will be responsible for all managerial aspects. They will confirm project benchmarks, schedule of completion, and deliverables; and ensure effective team communication. Working with the project Administrative Assistant, they will oversee doctoral student applications, internships placement, and curriculum coordination. A shared, secured document space will be used to manage internal working documents and the application process; the MRC website will host the LEADS-4-NDP doctoral application call, project news, and the curriculum. Blackboard will be used for lessons and assignments. PIs and advisory members will have quarterly meetings throughout the duration of the project, and minutes will be shared with project partners/mentors. Selected conferences, such as ALA, DPLA-fest, and CNI will also provide opportunities for team members to connect.

Budget and Resources. Total budget request to IMLS: $315,000.00, with a total budget of $466,822. The 1:1 cost share less student support costs will be met with a Drexel commitment of $153,553. Resources in terms of technical and software support, blackboard access, and MRC website are part of the university's basic infrastructure. Additional software and applications used are open source, and easily accessible. A budget justification covers details.

Advisory board members* will provide feedback to the program at critical points, give input on the curriculum design, and assess evaluation instruments, prior to implementation. **Partner/mentors** will contribute to the curriculum design, with real cases, and mentor the students during the 10-week internship period.

Advisory Board

- Devan Ray Donaldson, Assistant Professor, School of Informatics and Computing Indiana University, Bloomington
- Christopher Erdmann, Chief Strategist for Research Collaboration, NCSU Libraries
- Margaret L. Hedstrom, Professor, School of Information, University of Michigan
- Cliff Lynch, Executive Director, Coalition for Networked Information
- Richard Marciano, Director, Digital Curation Innovation Center, and Professor, College of Information Studies, University of Maryland
- Erik Mitchell, Associate University Librarian/Chief Information Officer, UC Berkeley, and Berkeley Institute for Data Science, Senior Fellow
- Antonios Zavaliangos, A. W. Grosvenor Professor of Materials Science and Engineering, Drexel University

*Letters of commitment are included as supporting documents.

2.4 Project Communication

A three-fold communication plan includes: 1) An all-project member communication plan (staff, LEADS Fellows, project mentors, and advisory board member), 2) Public dissemination plan, and 3) LEADS Fellow personal communication design plan. An all-project members mail list, will be used for communications among staff, LEADS Fellows, mentors, and advisory board members. A separate listserv will be set up for PIs and the Administrative Assistant for logistical communications. Advisory board members will be invited to join this listserv. The public dissemination plan will include quarterly news releases to update the broader LIS community on the LEADS-4-NDP program. Following year one of the project, outcomes and impact of the individual internship projects, as well as broader program news, will be highlighted. LEADS Fellows will be required to design and submit a communication plan for how they will communicate with their NDP mentor throughout the internship program. A special session at the bootcamp will address this requirement, and

review successful models used in the DataONE and RDA program. LEADS Fellows will work out their plan with the mentor, and submit the plan to the project instructors.

2.5 Evaluation and Outcome Assessment

LEADS-4-NDP has a two-fold evaluation plan focusing on: **1)** Student learning outcomes, drawing from IMLS's recommended Applied Research: Designing Evaluations (US GAO, 2012); and, **2**) Partner/mentor engagement success. We will use a multi-method approach and pursue evaluation during specific scheduled intervals. The timeline, methods, and goals are summarized below in Table 4.

Table 4: LEAD-4-NDP Evaluation Plan

Time period (*C1-18, C2-19)	LEADS-4-NDP Fellows (method/goal)	Project Mentors (method/goal)
May cohort acceptance	Profile survey. **Goal:** Assess data science experience and learning goals	Survey. **Goal:** Assess project communication and curriculum
During Drexel data science bootcamp	Focus group. **Goal:** Evaluate camp design	Focus group. **Goal:** Assess mentor engagement
Bootcamp completion	Survey **Goal:** Assess learning outcomes	N/A
Mid-way through summer internship	Semi-structured interviews (virtual or in-person) with a project PI. **Goal:** Assess learning outcome progress.	Survey. **Goal:** Mid-point assessment of LEAD-4-NDP Fellow
August internship completion	Survey. **Goal:** Assess individual and collective learning outcomes, and overall LEADS-4-NDP design	Survey. **Goal:** Assess LEADS-4-NDP program.
Fall 2019 only; LEADS-4-NDP completion	Focus group at a national conference (e.g., CNI), with LEADS Fellows and Mentors). **Goal:** Overall camp assessment and review of sustainability plans	

*C=Cycle. C1: May-Sept. 2018. C2: May-Sept. 2019.

The Evaluation cycle for C1 (the first group of LEADS Fellows), will inform the design for C2, specifically the recruitment, curriculum plans, and immersive internship design for the second cohort. Advisory board members will provide input to the test instruments. All evaluation will follow Drexel University's Institutional Review Board procedures for all evaluations in accordance with University policy, to ensure high standard and ethical behavior in our research approach. The survey component will specifically **measure the six curriculum learning outcomes** outlined in Section 2.2, Plans and Activities. The evaluation plan will be repeated during year two of the program, and conclude with the focus group at the national forum, engaging representatives from both cycles at a national conference.

2.6 Risk Amelioration

LEADS-4-NDP PIs recognize challenges associated with the proposed program. A planning meeting in December 2016, involving PIs, NDP partners, and advisory board members allocated time to discussing challenges, specifically the diversity of skills and knowledge that LEADS Fellows will bring to the table, and that there is no guarantee that all participants will become future faculty. To address the first problem, NDP partners will design their internship projects to scale at different levels. For example, there will be data science activities that can be addressed whether the student is competent with Excel or Python. Regarding future faculty, the LEADS-4-NDP application process requires applicants to share their future goals about becoming future faculty members. Applicants are also asked to commit to giving a guest lecture in a master's level class at their home institutions to share their project outcomes. This requirement will result in 18 lessons nation-

wide where fresh LIS/data science projects are brought to LIS master's curricula. While we cannot confirm that all 18 LEADS Fellows will become future faculty, it is likely that many will follow careers in academia, either as full-time faculty members or as adjuncts, so we are confident of LEADS-4-ND impact in bringing data science education to the next generation of LIS faculty and LIS education.

3. DIVERSITY PLAN

LEADS-4-NDP aligns with IMLS/LB21 goals and the NDP's need for a digital library workforce having an expanded skill set. The LEADS-4-NDP project includes a 3-pronged diversity plan for addressing these goals, covering **recruitment**, **outreach**, and **data science applications**: 1) **Recruitment:** We will work with the ALA Office for Diversity, Literacy & Outreach, our advisory board, project partners/mentors and colleagues at other institutions to reach students from diverse populations and seek their application to our program. A letter of support is provided by Gwendolyn Prellwitz, Assistant Director, ALA Office for Diversity, Literacy & Outreach Services, further outlining our plan in this area, 2) **Outreach**: LEADS-4-NDP will student showcase (virtual/in person) their work and share career goals roundtable at the Dornsife Center, West Philadelphia Obama Administration Promise Zone. PIs Greenberg and Lin work with the Dornsife Center, and have offered data science instruction to college bound high school girls, and presented the library profession. The showcase activity will allow LEADS Fellows experience in delivering education, and also allow us to reach diverse population of potential librarians. A letter of support from Katharine Travaline, Interim Director, Dornsife Center is included in supporting documents, 3) **Data science applications:** The NDP partnership—Mapping Inequality project, will allow development of data science lesson plans documenting minority issues. This NDP partnership presents a real-world opportunity for applying data science to NDP collections, raising the national dialog about race and diversity.

4. NATONAL IMPACT

LEADS-4-NDP aligns with IMLS/LB21 goals and the NDP's need for a digital library workforce having an expanded skill-set (IMLS, 2015). LEADS-4-NDP will advance doctoral student learning capacity, knowledge, and skill on a **national scale** via cross-training, hands-on experience, and collaboration. LEADS-4-NDP will have national impact in four key areas: **1)** LIS doctoral education enhancement, **2)** Nationwide doctoral student participation, **3)** NDP partner engagement from coast-to-coast, and **4)** Developing a vetted, cost-effective learning model of

LEADS-4-NDP will enhance LIS doctoral education in the data science area. A chief outcome of this project will be the development of a sustainable model. The PIs will summarize lessons learned and develop an educational model that is transferable and adaptable to different LIS programs. We will make our model, curriculum, and lesson plans accessible online to LIS programs nationally and will revise and publish the model and our experience in relevant LIS conference proceedings and journals. The LEADS-4-NDP program will extensible and transferable to other LIS programs. The LEADS-4-NDP will revise, finalize, and share the model toward the end of the program. The model will be shared via an open forum planned for the fall 2019 CNI conference. All LEADS Fellows, and LEAD-4-NDP partner-mentors, advisory board members, and core project staff (PI and administrative assistant) will share project outcomes and experience at relevant LIS conferences. Core project staff will also disseminate project outcomes and assessment results via the scientific and other scholarly literature.

Schedule of Completion

LIS Education And Data Science-4-the National Digital Platform (LEADS-4-NDP), Drexel University

YEAR →		17/18	2018			'18/19	2019			
Phases and Scheduled Tasks	Month →	N-J	F-A	My-Jl	Au-O	N-J	F-A	My-Jl	Au-0	S O
Phase 1 (*Nov. 2017-April 2018)*										
Prepare the LEADS-4-NDP curriculum modules, including case studies with sample partner data										
Refine LEADS internships with project partners										
Develop recruitment materials and application										
Seek feedback on curriculum and recruitment materials from partners and advisory board										
Distribute recruitment materials; open up the application process										
Review applicants; select and announce 2018 LEADS Fellows										
Phase 2 *(May 2018-Sept. 2018)*										
Invite LEADS fellows to Blackboard (online)										
Fellows complete 7-hour pre-camp training										
3-day bootcamp at Drexel University										
Fellows complete 10-week internship										
Phase 3 *(Oct. 2018-August 2019)*										
Review LEADS assessments; revise LEADS program for second iteration										
Seek feedback on program and curriculum revision from partners and advisory board										
Call for 2019 LEADS Fellows										
Review applicants; select and announce LEADS fellows										
Invite LEADS fellows to Blackboard (online)										
Fellows complete 7-hour pre-camp training										
3-day bootcamp at Drexel University										
Fellows complete 10-week internship										
Phase 4 *(Oct. 2018-Oct. 2019)*										
Draft sustainable model										
Feedback on sustainable model										
Dissemination of LEADS outputs and model										
Across the full project duration *(Nov. 2017-Oct. 2019)*										
Project assessment and evaluations[1]										
Advisory board meetings		During quarterly intervals								

[1] IMLS Performance Measure Statements: https://www.imls.gov/performance-measure-statements-and-information-learning-and-community-projects.

DIGITAL PRODUCT FORM

Introduction
The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (i.e., digital content, resources, assets, software, and datasets). The products 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 can be challenging. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, 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
You must provide answers to the questions in Part I. In addition, you must also complete at least one of the subsequent sections. If you intend to create or collect digital content, resources, or assets, complete Part II. If you intend to develop software, complete Part III. If you intend to create a dataset, complete Part IV.

PART I: Intellectual Property Rights and Permissions

A.1 What will be the intellectual property status of the digital products (content, resources, assets, software, or datasets) you intend to create? Who will hold the copyright(s)? How will you explain property rights and permissions to potential users (for example, by assigning a non-restrictive license such as BSD, GNU, MIT, or Creative Commons to the product)? Explain and justify your licensing selections.

Intellectual property rights: All the course materials will be published using CC0 (Creative Commons-Zero) in the iDEA repository at Drexel University.

Explanation for potential users: We will provide the CC0 indicator on these materials, with the indicator being hot-linked to the full description at MIT. Evaluation data gathered from the students and mentors will also be published using CC0, in Drexel's iDEA repository.

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms or conditions.

Ownership rights over products/conditions: With CC0, content developed for the curriculum in this project will be freely and globally accessible, as stated in the proposal. Drexel University will not claim any ownerships of the content, nor would it make any warranties about the work.

Justification: With this proposal, we seek to create and share educational resources. Open access would maximum the value of the investment and enhance the dissemination of results of this project.

Notification: Users will be notified by the CC0 indicator that will appear on the materials.

A.3 If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

Privacy: We will not create any curriculum products that have any privacy concerns. Evaluation data will not contain any Personally Identifiable Information (PII) data, and institutional names will be anonymized prior to data publication.

Part II: Projects Creating or Collecting Digital Content, Resources, or Assets

A. Creating or Collecting New Digital Content, Resources, or Assets

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

Digital content, resource, and assets created: This category include slides, lecture notes, and lessons plans, as well as example data and example codes for working with software.

Quantities/type/format: materials will cover the pre-camp online modules, "Preparation with project data and domain" and "Tutorials on data preprocessing" and the three-day boot camp. The amount of materials will cover approximately, 35 hours of lessons. Evaluation data will include survey data via Qualtrics, ported to excel, and WORD files, published as PDF of notes interviews and focus groups.

Resource types: include PDF files (notes), Power Point slides, CSV and Excel data files, Python and R code examples.

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

Office productivity Software: We will work with basic MS Office software, using Power Point, Excel, WORD and Notepad; Adobe Acrobat for publishing lecture notes.

Open-source resources: Python and R, two most commonly used programming languages for data science, will be introduced and discussed in the curriculum created in this project.

Open-access software: The curriculum will also include introductions and exercises on two most popular open-access data science software. One is **Tableau** from Tableau Software and the other is **Watson** from IBM. Both are commercial products with a public version for open-access.

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

File formats: pptx, docx, pdf, xlsx, twbx, py, pyc, txt, csv, .r, jpeg, tiff

B. Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).
All the digital content, resources, and assets created will go through the following reviewing and evaluation process:

1. The instructors will create first draft based on the proposed curriculum and time frames and do an initial review.
2. The PI/co-PIs will review all the materials and make sure they are consistent and connected.
3. The advisory board members are invited to review and comment on the teaching materials.
4. After the first summer camp, students and advisory board members will both complete evaluation surveys on the teaching materials and propose changes or updates.
5. PI/co-PIs/Instructors will collaborate to revise the teaching materials.
6. After the second summer camp, students and advisory board members will evaluate again the content and delivery of the teaching materials.
7. Final revisions will be done by PIs before posting the teaching materials to the University's iDEA repository.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance. Your plan may address storage systems, shared repositories, technical documentation, migration planning, and 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 C.F.R. § 200.461).

During the award period, items will be preserved on The College of Computing and Informatics' secure servers, which are monitored all the time for security and backed up regularly. For the long-term plan, items and resources will be preserved and maintained in Drexel University's iDEA repository, which is a trusted digital repository, and "guarantees future access to Drexel research and history through our commitment to preserve digital content through migration, secure backups and persistent URLs," (https://idea.library.drexel.edu/about).

C. Metadata

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

The iDEA repository supports Dublin Core and MODS metadata schemas. When items are deposited into the repository, all the items will be cataloged based on a custom MODS metadata schema, using controlled vocabulary from LCSH.

C.2 The iDEA repository's guarantees targeting preservation, secure backups, and migration over time, includes both content and associated metadata.

The iDEA repository's guarantees targeting preservation, secure backups, and migration over time, includes both content and associated metadata.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

The iDEA repository supports the exposure of digital contents via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Items to be added to the repository will be described based on a custom MODS metadata schema that facilitates discoverability and access.

A website, linked to the Metadata Research Center, will provide a summary description of the resources deposited in the iDEA repository (both curriculum materials and evaluation data). As part of the team's work to pursue scholarly and scientific disseminations via talks and publications, we will also highlight the open access of the resources, including access to the metadata used to describe these resources.

D. Access and Use

D.1 Describe how you will make the digital content, resources, or assets 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).

The digital content, resources, and other project materials will be made available to the public through two websites. One is the Drexel Metadata Research Center (MRC) home site: http://cci.drexel.edu/MRC/ . The other is the Drexel iDEA repository: https://idea.library.drexel.edu/ .

The MRC site will provide up to date materials and easy access interfaces for all the resources.

The iDEA repository will make the resources available on the Web as well as maintain a commitment for stability and long term preservation:

- Advanced search tools including facets and full-text searching
- A commitment of uptime, minimum 9x5, Monday-Friday 8:00am – 5:00pm.
- Exposure of digital contents via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
- Description of digital objects based on a custom MODS metadata schema that facilitates discoverability and access

Both websites are accessible via all standard web browsers. Special materials, such as Tableu files (twbx) or examples of python code (py, pyc) will require that users also have some form of software. In these cases, we will provide links to open access forms of this software.

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

- Example 1: Published visualization. Title: "Authors“Small Science” versus “Big Science”" https://idea.library.drexel.edu/islandora/object/idea%3A2846

- Example 2 Published lesson plan: Title: "Fast Food Frenzy" http://preview.tinyurl.com/zosoxst
- Example 3: Published data: Title: "Controlled Vocabularies for Scientific Data: Users and Desired Functionalities" https://figshare.com/articles/Controlled_Vocabularies_for_Scientific_Data_Users_and_Desired_Functionalities/1483478

Part III. Projects Developing Software

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) it will serve.

N/A

A.2 List other existing software that wholly or partially performs the same functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

N/A

B. Technical Information

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

N/A

B.2 Describe how the software you intend to create will extend or interoperate with relevant existing software.

N/A

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

N/A

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

N/A

B.5 Provide the name(s) and URL(s) for examples of any previous software 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 open-source licenses to maximize access and promote reuse. What ownership rights will your organization assert over the software you intend to create, and what conditions will you impose on its access and use? 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 and justify any prohibitive terms or conditions of use or access and detail how you will notify potential users about relevant terms and conditions.

N/A

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

N/A

C.3 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository:

URL:
N/A

Part IV: Projects Creating Datasets

A.1 Identify the type of data you plan to collect or generate, and the purpose or intended use to which you expect it to be put. Describe the method(s) you will use and the approximate dates or intervals at which you will collect or generate it.

Datasets to be created include evaluation data from survey, interview, and focus group. The data will be about the usefulness and effectiveness of the summer camp learning, data science curriculum, and internship projects. The data will be collected after each summer camp.

Format: Excel, WORD/notes, converted to PDF. Purpose: Project assessment and improvement.

A.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?

The evaluation protocol, including test instruments, will be reviewed by Drexel University's IRB prior to execution. This is stated in the proposal.

A.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).

No PII information will be collected, and no confidential or proprietary information will be collected.
Any institutional names associated with responders will be anonymized prior to public release.

A.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.

The College of Computing and Informatics (CCI) Drexel University has maintained secured servers for research and teaching. Consent forms and other information will be saved in a designated folder in the server that only the Pi/Co-PIs will have access to.

A.5 What methods 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).

Evaluation data: Survey data will be collected using Qualtrics software, and stored in Excel. Interview and focus group data will be captured in WORD and published as PDF documents. Basic technical requirements for Excel and PDF will be required to view or use the data.

A.6 What documentation (e.g., data documentation, codebooks) 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?

All data documentation, including codebooks with variable definitions and rating scales, will be published with the

datasets as a data package. Metadata descriptions will make this clear. PI Prof. Greenberg has extensive experience in this area, overseeing Dryad repository metadata for 8 years.

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

We will deposit all data, including documentation, in the iDEA respository at Drexel University. As noted, The iDEA repository "guarantees future access to Drexel research and history through our commitment to preserve digital content through migration, secure backups and persistent URLs," (https://idea.library.drexel.edu/about). We will also describe the data and provide URLs to the data in our presentations and publications.

A.8 Identify where you will deposit the dataset(s):

Name of repository: iDEA Repository at Drexel University

URL: https://idea.library.drexel.edu/

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

We will review the data management plan at the end of phrase 1 and phase 2, Sept.-Dec. 2018. We will make necessary modification at the beginning of Phase 3, Jan.-April, 2019.

We will make quarterly manual checks on our secure servers, to make sure data is safe. Drexel University's IT support group also performs weekly back-ups of the servers. At the close of the project, we will conduct a final review before all the data are published.