Drexel University and partners request $887,153.97 for the **LIS Education And Data Science-Integrated Network Group (LEADING)**, a Laura Bush 21st Century Librarian (LB21) National Digital Infrastructures and Initiatives project. The LEADING project scales-up the highly successful LEADS initiative by extending the data science educational pipeline to a much broader range of participants. LEADING will prepare a diverse, nation-wide cohort of **50 LIS early career librarians and doctoral students** for data science endeavors. LEADING's model includes community hubs (Montana State University, University of California San Diego, and OCLC), along with 14 member nodes, serving as mentoring sites (see Figure 1). LEADING Fellows, community hubs, and member nodes ultimately will form a **network that will advance and catalyze data science throughout our national digital infrastructure**.

1. STATEMENT OF BROAD NEED

Over the last few years, there has been a rapid increase of data science efforts to meet the growing demand and opportunity for data-related expertise across the library environment (Palmer, et al., 2018). One such program has been the **LIS Educational and Data Science (LEADS)** program, led by Drexel University. LEADS prepared 21 LIS Fellows in data science, leveraging Drexel University's formal Master's in Data Science (MSDS) curriculum, and through collaboration with leading national digital infrastructure sites (e.g., California Digital Library, Digital Scholarship Center-Temple University, OCLC, and Smithsonian Libraries, among others). LEADS Fellows participated in both online and in-person formal curricula, and undertook an immersive data science experience, working with a mentor (or mentors) at one of the partner mentoring sites. The outcomes of this initiative are quite remarkable, including:

- Extensive output in the form of papers, posters, and presentations (see **Suppportingdoc5** for a complete list)
- Ongoing inquiries from libraries seeking to hire LEADS Fellows (most recently, the National Institute of Environmental Health Sciences Library reached out to hire a LEADS Fellow)
- Continued interest from both front-line professionals and LIS doctoral students in the next opportunity to participate in LEADS

The LEADING project is motivated by these outcomes, particularly the overwhelming interest of many front-line early career professionals and LIS doctoral students, who express their need for formal data science training, and who are eager to learn and gain experience. The LEADING proposal specifically targets these needs, and will **extend formal educational efforts beyond academic programs, building a cohesive network engaging the spectrum of LIS data science stakeholders.** The value of this approach is made clear by observing current data science developments across *LIS education*, *professional training*, and *library user services.*

• **Formal education**. Educators have investigated data science curricula (Ortiz-Repiso, et al., 2018; Han & Zu, 2017; Song & Zhu, 2016), and several iSchools have recently launched formal data science graduate degrees, Drexel University among the first. Initiatives have also targeted identifying core competencies and data science skills (Burton, et al., 2018); and, in early 2019, the iCaucus formed the iSchool Model Data Science Curriculum Committee, which is charged developing a standard iSchool data science curriculum distinct from computer science or business-oriented data science curricula (Oh, et al., 2019). With the exception of post-graduate certificates, these efforts are at early stages, and have not had time to mature and cross-over to support front-line professional training needs.

• **Professional Training**. Front-line professional training opportunities have emerged through initiatives such as Collections as Data, promoting computational use of collections as data, and data science approaches; the Library Carpentry, providing hands-on training for information professionals; and LC Labs, where information professionals are, in essence, residential data scientists, with an opportunity to learn by working with the Library of Congress' data resources. Advocacy for identifying key skills and building in-house talent is made clear in two recent, significant reports by Federer, et al. (2020) and Padilla (2019). While important, there is a notable absence in these analyses of a link to formal education, although LIS educators were consulted while developing these reports. Further, LIS programs have not had sufficient time to study and fully connect with these exemplary initiatives.

• **User Service**. Finally, libraries have been advancing user data services. Key examples include the University of Arizona Libraries program focusing on data science literacy (Oliver, et al., 2019) for library users across various disciplines, and North Carolina State University's well-known Libraries Data And Visualization workshop series. These initiatives rely on information professionals with data science expertise, who are in high demand.

These developments are significant, demonstrating essential components of the LIS data science ecosystem; and simultaneously making apparent the following challenges:

1. Fragmentation of data science educational opportunities across LIS.
2. Absence of a cohesive, data science network bringing together all stakeholders (educators, front-line professionals, students, administrators, and researchers).

This fragmentation and absence of a cohesive network are chiefly due to the acceleration of data production, and the pressing need for data science expertise across all fronts. These challenges highlight the need to extend formal educational efforts beyond academic programs, and build a more cohesive network. We address these needs with LEADING, drawing from the success of LEADS, building on lessons learned and developing a LIS data science network supporting the exchange of ideas, resources, and strategies.

2. PROJECT DESIGN

We propose the LEADING project, creating a **national network of early career professionals, LIS doctoral students**, and other stakeholders who will collectively advance LIS capacities in data science and substantially enhance the growth and sustainability of our national digital infrastructure.

2.1 Project Goals

LEADS is designed to address the following four goals:

1. Extend the Education Pipeline. Drexel University's iSchool is one of the first LIS programs to implement data science curricula at the undergraduate, master's, and doctoral levels. The LEADING project will leverage Drexel's curriculum and faculty expertise and extend the educational pipeline to many more LIS doctoral students and early-career frontline information professionals.

2. Collaboration and Community Development. LEADING will bring together educators, researchers, library leaders, frontline information professionals, and doctoral students, who will collaborate and advance data science in LIS education and practice. The cohort of LEADING Fellows will develop invaluable, enduring partnerships supporting future collaboration.

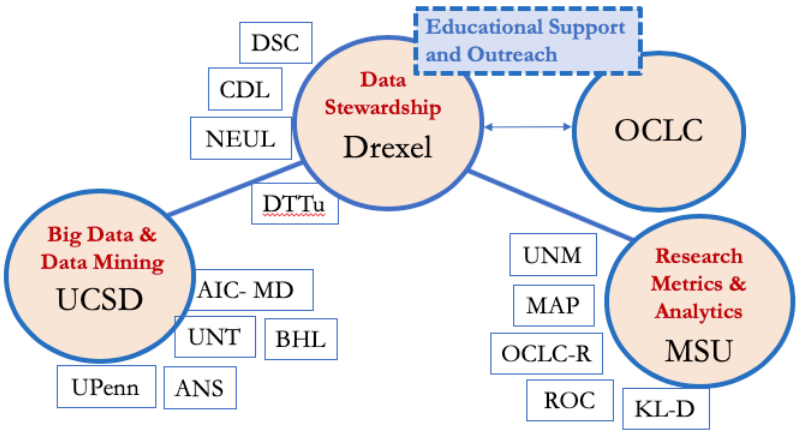
3. Diversity and Inclusivity. During the original LEADS project, over half the participants represented racially diverse communities, and almost 30% were from under-represented minorities. The LEADING project will carry this goal forward through our diversity plan, supported by a **Diversity and Inclusivity Task Force** of previous LEADS Fellows.

4. Sustainability. A foundation for a LEADS sustainability model emerged with LEADS, as both doctoral students and early career individuals inquired about participation, even without a stipend. Additionally, Temple University and Montana State University each joined LEADS, offering to support Fellows. These occurrences offer proof of, and suggest a direction for, developing LEADING as a sustainable network, accommodating those who have resources, as well as those who do not.

2.2 LEADING'S Model

LEADING's project design adapts the National Science Foundation's (NSF) Big Data Innovation Regional Hub model, and is comprised of a set of partner hubs and member nodes, with the capacity to grow. The LEADING model (Figure 1) includes Drexel University as the Central Coordinating and Educational Hub; Montana State University (MSU) and University of California, San Diego (UCSD) as Community Hubs; and OCLC as a Co-Educational Hub. Member Nodes are clustered around Drexel University, MSU, and UCSD based on themes reflecting the data science Fellowship projects member nodes.

***Figure 1*: LEADING's Hub Model: Themes and Member Nodes**



Member Nodes

- Academy of Natural Sciences (ANS)
- AI-Collaboratory, U. of MD iSchool (IAC–MD)
- California Digital Library (CDL)
- Digital Scholarship Centr., Temple U./ (DSC)
- Digital Scholarship, Tish Library, Tufts University (DTTu)
- Kislak Center for Special Collections, Univ. of Pennsylvania Libraries (UPenn)
- Kline Law Library-Drexel (KL-D)
- Movement Alliance Project (MAP)
- Northeastern University Libraries (NEUL)
- OCLC—R&D (OCLC-R)
- Smithsonian Libraries, Biodiversity Heritage Library (BHL)
- University of New Mexico (UNM)
- University of North Texas (UNT)
- University of Rochester (ROC)

The clustering approach illustrated above will also support project management and community building. An important take-away from LEADS was that regular, virtual check-ins and faculty/Fellow interaction over the internship period was integral to project success. The network model of hubs and nodes will enable LEADING to effectively continue this project component. Additional modifications to LEADING, drawing from LEADS and outcomes (Greenberg, et, al., 2020) include the following:

- LEADING Fellowship extends the original 10-week immersive experience to a 6-month fellowship period. Nearly all the 21 LEADS Fellows continued their work beyond the 10-week immersive Fellowship period; and several 2019 Fellows are still active in LEADS scholarship. In fact, one LEADS Fellow's immersive experience has become their dissertation topic.

- Community Hubs, MSU and UCSD, along with Drexel University will also serve as venues for sharing LEADING's curriculum and for implementing the sustainability plan during project year-3, through a local Community Hub outreach workshop, led by LEADING project PIs and Fellows.
- OCLC, as a Co-Educational Hub, will sponsor two data science challenges (project year-1 and year-2), in which LEADING Fellows, as well as other LIS students and information professionals will gather for two days to work with OCLC data and develop curriculum materials to be shared via OCLC and LEADING's GitHub portal, and accessible to both LIS data science programs as well as front-line professionals

LEADING Fellowship: Application and Evaluation/Selection

The program will provide stipends for 50 early- to mid-career LIS professionals and LIS doctoral students from ALA accredited programs across the country over the three year initiative. Community hubs will have flexibility in recruiting Fellows. The LEADING Fellowship application and selection process will be modeled after the successful approach implemented for the two LEADS cohorts, with a few modifications to accommodate early career professionals and their applications (Table 1 gives provides an overview of criteria and process).

TABLE 1: LEADING Fellowship: Application and Evaluation/Selection

Leading Fellowship Application
LEADING application process will be made through secure server. Applicant requirements: **1)** statement of interest identifying desired learning goals, **2)** CV, **3)** letter of reference from immediate supervisor (for early career members) or academic advisors (for doctoral students), and **4)** completion of a questionnaire to assess applicant knowledge of skills specific to coding (R, Python, Java, HTML, etc.) and statistical software packages (e.g., Excel, SAS, SPSS, Tableau), and will assist LEADING faculty in updating and delivery of the curriculum. The applicants will also be asked to select their top three Fellowship sites. Additionally, all Fellows will be asked if they need stipend support, travel support to the boot camp, or both. Based on preliminary analysis, we anticipate that some, but not all, early career professionals may forego the stipend, due to in-house support. In this case, unused support will be allocated toward year-3, to another early-career professional, and assist with testing the sustainability model.
Evaluation/Selection Process
Member Node leaders will rank their top three preferred candidates, based upon the interests and expertise of the candidate in relation to the proposed fellowship project and the applicant's learning goals. Guided by these rankings and overall criteria, LEADING PIs will recommend candidates for each Member Node project, who may accept or ask to reassess. Community hubs MSU and UCSD will have the flexibility to select among the LEADING applicant pool and their internal employees. The selection process will aim for an even balance of early-career professionals and doctoral students. While most member nodes will host only one LEADING Fellow, there will be a few cases in which two Fellows will collaborate. This approach was tested with LEADS in three instances, and was quite successful.

2.3 Drexel Setting and the LEADS Curriculum

All LEADING fellows will complete: **1)** a 15-hour online preparatory curriculum, **2)** an intensive 4-day data science bootcamp, at Drexel University, and **3)** a 6-month data science internship with a LEADING Member Node.

The LEADING curriculum builds off of Drexel's strengths, positioned in the College of Computing and Informatics (CCI). Drexel's Information Science (IS) Department is a pioneering institution in research-oriented LIS education. Drexel's iSchool in particular has set numerous precedents in the community, including early development of

fully-online programmatic offering of the ALA accredited MSLIS, and other degrees across the University. The curriculum, outlined below, is based on the initial delivery of LEADS, with modifications drawing from LEADS feedback and experience, along with new developments in Drexel's data science program, including new offerings focusing on artificial intelligence, data preservation, and human centered data science.

Curriculum

LEADING's curriculum and lesson plans will be made openly accessible to the LIS program community, both nationally and internationally, through the Metadata Research Center LEADS Github Page, where the LEADS curriculum is currently shared. The curriculum for a 4-day boot camp follows major competencies within the EDISON model curriculum (Demchenko, et al., 2016; Wiktorski, 2017), and further incorporates work on computational thinking associated with the computational archives community (Marciano, et al., 2019, Underwood, et al., 2018). One of the leaders in this area, Richard Marciano, is on the LEADING Advisory Board, and delivered a module on this topic at the 2019 LEADS boot camp. Table 2 and 3 provide a high-level overview of the curriculum, and a more detailed outline is found in proposal **Suppportingdoc6**.

Table 2: Online Curriculum

Online modules before the boot 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 3: LEADING Boot Camp Curriculum

Day 1: Data and Knowledge	**Day 3: Computing Skills and Engineering**
Big Data and Data Science	**Data Mining and Machine Learning tools**
Big Data and Data Science: the Fourth Paradigm Data Science processes Data science projects in LIS	Text Processing, Mining, Scalability, Common Matrix Factorization Processes and Time/Space Complexity Parallelization
Data Management	**Automated Data Analysis Tools**
Data sources and types, Relational database, analytics w/SQL, Large-scale data storage	Review of automated data analytics tools, Watson Analytics, BigML, etc., Artificial intelligence (AI)
Data Integration and Quality	**Large-scale Data Intensive** Computing
Metadata, usability, and data quality Big Metadata, Ontologies Linked data for data science Data preservation	Cloud computing for big data MapReduce, Hadoop, Spark Parallel framework, graph computing
Day 2: Methods for Analytics	**Day 4: Evaluation and Presentation**
Probability, Statistical methods, Machine learning approaches Preprocessing, Cleansing Unsupervised and Supervised learning Graphical models, neural networks	Evaluation procedure, metrics, and use Visual analytics and information visualization Presentation and communication Human centered data science

The LEADING boot camp has been extended from two-and-a-half days, which was the LEADS practice, to four days, given feedback from the LEADS Forum. More time will enable group work, and further support community building.

• **Internships**
A signature component of the original LEADS program was the immersive internships with national digital infrastructure partners. LEADING will continue this aspect, working with member nodes, and extending the immersive learning experience to a 6-month period. Each LEADING Fellow will address questions and work with a mentor and a faculty member. **Supportingdoc3.pdf** includes a list of mentors, their institutions, and letters that identify the collection that will be used to shape the Fellowship. LEADING Faculty will work with member nodes to design the data science Fellowship opportunities, which will be posted in advance of the application process. LEADING faculty and member node mentors will work with an updated version of the LEADS Template, which captures essential information about each project's design, goals, and learning outcomes. Fellows will be matched to these data science projects according to their skill, experience, and selected areas for growth.

2.4 Phases and Activities
LEADING consists of the following four phases:

• **Phase 1: Curriculum Enhancement and Recruitment** *(Nov. 2020-April 2021).* Project PIs prepare curriculum modules covering introduction to data science, data curation, big metadata, algorithms, analytics, and visualization. Curriculum enhancements will include AI (artificial intelligence) and data preservation. Hub partners will contribute to curriculum development. Phase 1 will also include virtual meetings with Hub partners and NDP mentors to design Fellowship projects and initiate recruitment.

• **Phase 2: LEADING Fellows/Project Launch** *(May 2021-Aug. 2021).* Fellows will complete the curriculum outlined below, including a 6- month immersive data science internship with their Member Node institution or at their Community Hub. Internships will be virtual or on-site. During Phase 2 we will conduct formal assessments within the IMLS evaluation framework. Our approach will consist of surveys, interviews, and focus groups with students, hub partners, Member Node project mentors, and faculty discussions with the Advisory Board.

• **Phase 3: LEADING Project 2** *(Sept. 2021-Aug. 2022).* The project team will review and revise the LEADING Fellowship program for its second iteration. Phases 1 and 2 will repeat, and Fellows participating in both phases will disseminate results and learning experiences at relevant venues.

• **Phase 4: Sustainability** *(Sept. 2022-Oct. 2023).* During Year 3, we will gather input from Community Hub partners, Member Nodes, mentors, advisory board members, and Fellows, to develop and test the sustainability plan. Feedback from LEADS has been instrumental here, and the aim for year-3 is to engage at least eight LEADING Fellows, who will work with a mentors who have been a part of LEADING, or who the Fellow applicant has identified. The sustainability model will not provide Fellows stipends, but will offer support to attend the Data Science boot camp at Drexel, for those demonstrating need. LEADING communication protocols with regular check-ins will be followed, and Fellows will form a cohort, to share their experiences. Drexel LEADING faculty are committed to testing this model, and see opportunity to engage more than eight Fellows, using

Drexel's online facilities. Details of the sustainability plan will be worked out further in coordination with all LEADING members, and guidance from LEADING's Advisory Board.

2.5 Personnel and Management

Faculty and Staff: LEADS includes seven faculty members from the Information Science Department, College of Computing and Informatics, Drexel University, and a Project Manager. Jane Greenberg, Alice B. Kroeger Professor, and Director of the Metadata Research Center (MRC), will serve as the Project Director and a LEADING instructor. Co PIs, Il-Yeol Song, Professor; Weimao Ke, Associate Professor; Erjia Yan, Associate Professor; and Jake Williams, Assistant Professor, have all been a part of LEADS, and will serve as LEADING curriculum instructors. Additional faculty members joining LEADING include Teaching Professor Lei Wang, and Assistant Professors Alex Poole and Mat Kelly, all of whom have joined the initiative to ensure appropriate mentoring, and to extend to the new topics: artificial intelligence, data preservation, and human centered data science. The core project staff also includes a project Program Manager, who will work closely with the project PIs to coordinate project meetings, connections with all project hubs and nodes, organize the Fellows application process, and assist with organizing the summer boot camp (see **Resumes.pdf**).

Co-Educational Hub Leaders: Andrew Pace, Executive Director for Technical Research, will serves as the Co-Educational Hub leaders and will collaborate with LEADING team members in the development of a "data science challenge" for LEADING Fellows, LIS students and information professionals, generating excitement about OCLC data and learning about data science (see **Resumes.pdf**).

Community Hub Leaders: Community Hub leaders, Erik Mitchell, University Librarian, University of California San Diego; and Kenning Arlitsch, Dean, Montana State University Library, will engage selected staff to participate in LEADING and serve as a link between designated Drexel faculty and affiliated Member Node mentors and Fellows (see **Resumes.pdf**).

Mentors: Member Node mentors will contribute to the curriculum design, provide data use cases for the curriculum and access to data for their designated Fellows, and mentor the fellows during the 6-month internship. Project mentors represent leading institutions with data services and collections that are ideal sites for LEADING Fellows. Figure 1, above, lists the member nodes. The mentors are listed in the **Projectstaff.pdf** attachment, and letters of support are included in **Supportingdoc3**.

Advisory Groups (see Table 4)

- **Leading Advisory Board:** Advisory board members will provide feedback on the curriculum design, interact with Fellows and mentors, and give input and guidance on the development of the sustainability plan.

- **Diversity and Inclusivity Taskforce:** Taskforce members will provide feedback on the diversity and inclusivity considerations with respect to curriculum design, recruitment plans, and sustainability plans.

Table 4: LEADING Advisory Groups

LEADING Advisory Board	
▪ **Devan Ray Donaldson**, Assistant Professor, Department of Information and Library Science (ILS), School of Informatics, University of Indiana ▪ **Christopher C. Erdmann**, Engagement, Support, and Training Expert, NHLBI BioData Catalyst, RENCI, University of North Carolina at Chapel Hill ▪ **Clifford Lynch**, Executive Director, Coalition for Networked Information (CNI) ▪ **Richard Marciano**, Professor and Director, AI-Collaboratory, University of Maryland iSchool	▪ **Mark Phillips**, Associate Dean for Digital Libraries, University of North Texas Libraries (LEADS Fellow 2018) ▪ **Juliane Schneider**, Team Lead, eagle-I, Harvard Catalyst, Clinical and Translational Science Center ▪ **Megan Finn Senseney**, Head, Office of Digital Innovation and Stewardship, University of Arizona Libraries
LEADING Diversity and Inclusivity Task Force	
▪ **Yi Yun (Jessica) Cheng**, 2019 LEADS Fellow, PhD Student, University of Illinois at Urbana–Champaign ▪ **Sonia Pascua**, 2019 LEADS Fellow, PhD Student, Drexel University ▪ **Hanlin Zhang**, 2019 LEADS Fellow, PhD Student, UNC Chapel Hill	▪ **Gretchen Stahlman**, 2018 LEADS Fellow, Assistant Professor, Rutgers University ▪ **Adam Johs**, 2018 LEADS Fellow, PhD Student/ Drexel University

Budget and Resources

The total budget request to IMLS is $887,153.97; and the project's full budget is: $1,308,399.64. The 1:1 cost share less student support costs will be met with a Drexel commitment of $421,245.67. Resources in terms of technical and software support, blackboard access, and the MRC website are part of Drexel University's basic infrastructure. Additional software and applications are open source and easily accessible. Drexel has first-rate in-house technical expertise to support the LEADING project and virtual communication software. A budget justification covers details.

Communication and Collaboration

LEADING's communication plan includes: 1) A project leader communication plan (staff, LEADING PIs, Community Hub leaders, project mentors, and advisory board members), 2) Public dissemination plan, 3) Community Hub communication plan, and 4) LEADING fellow personal communication plan. A project leader mailing list will be used for communications among project leaders across Drexel, Community Hubs, Member Nodes, and advisory board members. The public dissemination plan will include quarterly news releases to update the broader LIS community on the LEADING program. During Year-3 of the project, outcomes and impact of the individual internship projects, as well as broader program news, will be highlighted. Drexel's LEADING PIs will work with Community Hub leaders to schedule regular virtual check-ins over the 6-month fellowship period. LEADING Fellows will be required to design and submit a communication plan for how they will communicate with their Member Node and Community Hub mentors throughout the internship program. A special session at the bootcamp will address this requirement, and review the successful models used by our previous LEADS-4-NDP models, as well as models used in the DataONE and RDA programs. LEADING fellows will develop their plans with their mentors, and then submit them to the project instructors.

Evaluation

LEADING's evaluation plan will address: 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, as pursued with LEADS, performing evaluations at scheduled intervals. The timeline, methods, and goals are summarized below in Table 5:

Table 5: LEADING Evaluation Plan

Time Period (yearly)	LEADING Fellows (method/goal)	Project Mentors (method/goal)
April - post-cohort acceptance	Profile survey. **Goal:** Assess data science experience and learning goals	Survey. **Goal:** Assess project communication and curriculum
June - during-Drexel data science boot camp	Focus group. **Goal:** Evaluate camp design	Focus group. **Goal:** Assess mentor engagement
September - post-Boot camp completion	Survey **Goal:** Assess learning outcomes	Survey. Goal: Obtain mentor feedback on Fellow/s
September - Mid-way through 6-month internship	Semi-structured interviews (virtual or in-person) with Member Node leaders and Community Hub Leaders. **Goal:** Assess overall experience and learning outcome progress.	Survey. **Goal:** Mid-point assessment of LEADING Fellow
December - internship completion	Survey. **Goal:** Assess individual and collective learning outcomes, and overall LEADING design	Survey. **Goal:** Assess LEADING program.
Year-3, December	Focus group at a national conference (potentially CNI) with LEADING fellows, Community Hub leaders, and Member Node leaders. **Goal:** Overall assessment and review of sustainability plans.	

Risk Mitigation

LEADING PIs recognize challenges associated with the proposed program. Several planning meetings have taken place over the last few months involving PIs, Community Hub Leaders, OCLC representatives, Member Node partners, and advisory board members. Additionally, LEADS Fellows gave input at the recent LEADS Forum. The most recent challenge has been the threat of Covid-19, and other concerns include the size of LEADING, and risks associated with scaling up a highly successful program. These two noted risks can be addressed as follows. LEADS and the proposed LEADING is extensively virtual, as the Fellows have for the most part been connected to mentoring sites virtually. The team is confident that with Drexel's infrastructure, the LEADING program, including the bootcamp, can be run virtually if necessary. The challenge will be in cohort building, which has been a trademark of LEADS. Faculty liaisons to each of the community hubs will take on extra responsibility to build the cohort, should the LEADING program need to be offered virtually. The LEADING model, with Community Hubs and Member nodes has been designed specifically to address the issue of size, and this approach will be assessed and modified as necessary throughout the program. The Community Hub leaders will

meet with LEADS PIs, and provide feedback, and the hubs can learn from each other. Member Node partners will design their internship projects to scale at appropriate levels, as necessary. A final risk considered with LEADS was potential limitations in the level of data science activities that Fellows would address. This concern was successfully addressed by LEADS by requiring that applicants have at least some comfort with Excel and demonstrate a willingness to learn. This requirement will be continued with LEADING, and we anticipate this concern to be a low risk, given the experience with LEADS.

3. DIVERSITY PLAN

LEADING's diversity plan covers 1) recruitment, 2) LEADING curricula, and 3) Fellowship opportunities. Recruitment: We will work through the ALA Office for Diversity, Literacy, and Outreach Services to recruit LEADING Fellows under-represented populations. Drexel and hub partners will also work through their corresponding Offices of Equality and Diversity to ensure diversity and inclusivity in recruiting early career professionals. Collections: Fellows will have access to unparalleled collections via Movement Alliance Project for Social Justice; Northeastern University's collections document diversity, with historical records of Boston's African American, Asian American, LGBTQ, and Latino communities; the AIC's Collaboratory Network Mapping Inequality project, as well as the Densho.org project with Japanese internment camp data. LEADING faculty will draw from these collections in delivery of the curricula, and in helping Fellows scope insightful questions; in addition, several LEADING Fellows will be able to pursue their 6-month immersive internship working with these selected sites. Finally, as noted above, LEADING has formed a Diversity and Inclusivity Taskforce to provide feedback on inclusivity considerations of the curriculum, recruitment plans, and sustainability plans. The Taskforce is comprised of previous LEADS Fellows.

4. BROAD IMPACT

LEADING will have broad, national impact in three key areas: First, LEADING will leverage Drexel's curriculum and faculty expertise to extend the educational pipeline to many LIS doctoral students and early-career frontline information professionals from institutions throughout the country. This breadth is a key component of the project's impact. Second, LEADING will facilitate Collaboration and Community Development, bringing together educators, researchers, library leaders, frontline information professionals, and doctoral students, who will collaborate and advance data science in LIS education and practice. The cohort of Fellows is an integral part of developing a network, and LEADING will provide the infrastructure to create invaluable, enduring partnerships supporting future collaboration. Third, LEADING will develop and test a sustainability model supporting growth of a data science network. The network model will help ameliorate the current fragmentation across LIS education, professional training, and user-services, building a cohesive community supporting the exchange of ideas, resources, and strategies. The network including all stakeholders, with a focus on data science, is also critical, as libraries also begin to grapple with artificial intelligence (AI) (ADIR, 2019; Stanford Libraries, 2019). Ultimately, the LEADING network will advance and catalyze data science by engaging stake all holders across our national digital infrastructure.

Schedule of Completion for LEADING

Year 1-Year 2												
Quarter ->	1	2	3	4	1	2	3	4	1	2	3	4
Phases and Scheduled Tasks												
Phase 1: Curriculum Enhancement and Recruitment (Nov. 2020-April 2021)												
Project PIs prepare curriculum modules covering introduction to data science, data curation, big metadata, algorithms, analytics, and visualization.												
Virtual meetings with Community Hub partners and member nodes to design Fellowship projects and initiate recruitment.												
Develop recruitment materials and application												
Seek feedback on curriculum and recruitment materials from Community Hubs, Member Nodes, and Advisory Board												
Distribute recruitment materials; open up the application process												
Review applicants; select and announce 2021 LEADS Fellows												
Phase 2: LEADING Fellows/Program Launch (May 2021-Aug. 2021)												
Invite LEADS fellows to Blackboard (online)												
Fellows complete 15-hour online curriculum												
3-day data science bootcamp at Drexel University												
Fellows meet Project mentors at Member Nodes or Community Hubs in person or virtually												
Fellows complete a 6 month immersive data science internship with their member node or at their Community Hub, virtually or on site.												
OCLC Data Challenge Meeting at OCLC headquarters												
Conduct formal assessments within the IMLS evaluation framework												

Year 3 - Phase 3: LEADING Program 2 (Sept/Oct. 2021-Aug. 2022)												
Review and revise the LEADING Fellowship program for its second iteration												
Seek feedback on program and curriculum revision from Community Hubs, Member Nodes, and Advisory Board												
Call for 2022 Fellows												
Review applicants; select and announce 2022 LEADING fellows												
Invite LEADING fellows to Blackboard (online)												
Fellows complete 15-hour online curriculum												
3-day data science bootcamp at Drexel University												
Fellows meet Project mentors at Member Nodes or Community Hubs in person or virtually												
Fellows complete a 6 month immersive data science internship with their member node or at their Community Hub, virtually or on site.												
OCLC Data Challenge Meeting at OCLC headquarters												
Phase 4: Sustainability (Sept. 2022-Oct. 2023)												
Two Community Hub outreach workshops												
Gather mentors, Community Hub and Member Node partners, advisory board members, and Fellows' inputs, to develop a sustainability plan.												
Draft sustainability plan												
Dissemination of LEADING outputs												

DIGITAL PRODUCT FORM

INTRODUCTION

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to digital products that are created using federal funds. This includes (1) digitized and born-digital content, resources, or assets; (2) software; and (3) research data (see below for more specific examples). Excluded are preliminary analyses, drafts of papers, plans for future research, peer-review assessments, and communications with colleagues.

The digital products you create with IMLS funding require effective stewardship to protect and enhance their value, and they should be freely and readily available for use and reuse by libraries, archives, museums, and the public. 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

If you propose to create digital products in the course of your IMLS-funded project, you must first provide answers to the questions in **SECTION I: INTELLECTUAL PROPERTY RIGHTS AND PERMISSIONS.** Then consider which of the following types of digital products you will create in your project, and complete each section of the form that is applicable.

SECTION II: DIGITAL CONTENT, RESOURCES, OR ASSETS

Complete this section if your project will create digital content, resources, or assets. These include both digitized and born-digital products created by individuals, project teams, or through community gatherings during your project. Examples include, but are not limited to, still images, audio files, moving images, microfilm, object inventories, object catalogs, artworks, books, posters, curricula, field books, maps, notebooks, scientific labels, metadata schema, charts, tables, drawings, workflows, and teacher toolkits. Your project may involve making these materials available through public or access-controlled websites, kiosks, or live or recorded programs.

SECTION III: SOFTWARE

Complete this section if your project will create software, including any source code, algorithms, applications, and digital tools plus the accompanying documentation created by you during your project.

SECTION IV: RESEARCH DATA

Complete this section if your project will create research data, including recorded factual information and supporting documentation, commonly accepted as relevant to validating research findings and to supporting scholarly publications.

SECTION I: INTELLECTUAL PROPERTY RIGHTS AND PERMISSIONS

A.1 We expect applicants seeking federal funds for developing or creating digital products to release these files under open-source licenses to maximize access and promote reuse. What will be the intellectual property status of the digital products (i.e., digital content, resources, or assets; software; research data) you intend to create? What ownership rights will your organization assert over the files you intend to create, and what conditions will you impose on their access and use? Who will hold the copyright(s)? Explain and justify your licensing selections. Identify and explain the license under which you will release the files (e.g., a non-restrictive license such as BSD, GNU, MIT, Creative Commons licenses; RightsStatements.org statements). 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.

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. Aggregated, anonymized valuation 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 among the hub (Drexel), co-ed hub (OCLC), community hubs, and member nodes. 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 will be used for in-house assessment only, and institutional names will be anonymized.

SECTION II: 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 the format(s) 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 online and f2f learning modules, "Preparation with project data and domain" and "Tutorials on data preprocessing", four-day boot camp, and six-month internship. The amount of camp related materials will cover approximately 35 hours of lessons. Evaluation data will include survey data via Qualtrics, ported to excel, and WORD files.
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 digital 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 popular open-access data science software. One is Tableau Academics from Tableau Software and the other is Watson from IBM. Both are commercial products with a free academic version for open access.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG, OBJ, DOC, PDF) you plan to use. If digitizing content, describe the quality standards (e.g., resolution, sampling rate, pixel dimensions) you will use for the files you will create.

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

Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan. How will you 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. Community hub coordinators 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.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period. Your plan should 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).

Metadata

C.1 Describe how you will produce any and all technical, descriptive, administrative, or preservation metadata or linked data. Specify which standards or data models you will use for the metadata structure (e.g., RDF, BIBFRAME, 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 Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.

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.

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, delivery enabled by IIIF specifications).

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

D.2. Provide the name(s) and URL(s) (Universal Resource Locator), DOI (Digital Object Identifier), or other persistent identifier 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: Citation sentiment data https://doi.org/10.6084/m9.figshare.7837517.v2
Example 4: 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

SECTION III: SOFTWARE

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 or similar functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

N/A

Technical Information

B.1 List the programming languages, platforms, frameworks, 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), URL(s), and/or code repository locations for examples of any previous software your organization has created.

N/A

Access and Use

C.1 Describe how you will make the software and source code available to the public and/or its intended users.

N/A

C.2 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository:

N/A

URL:

N/A

SECTION IV: RESEARCH DATA

As part of the federal government's commitment to increase access to federally funded research data, Section IV represents the Data Management Plan (DMP) for research proposals and should reflect data management, dissemination, and preservation best practices in the applicant's area of research appropriate to the data that the project will generate.

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

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 used for in-house analysis only.

A.3 Will you collect any sensitive information? This may include personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information. If so, detail the specific steps you will take to protect the information while you prepare it for public release (e.g., anonymizing individual identifiers, data aggregation). If the data will not be released publicly, explain why the data cannot be shared due to the protection of privacy, confidentiality, security, intellectual property, and other rights or requirements.

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

A.4 What technical (hardware and/or software) requirements or dependencies would be necessary for understanding retrieving, displaying, processing, or otherwise reusing the data?

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.5 What documentation (e.g., consent agreements, data documentation, codebooks, metadata, and analytical and procedural information) will you capture or create along with the data? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the data it describes to enable future reuse?

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 10 years.

A.6 What is your plan for managing, disseminating, and preserving 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.7 Identify where you will deposit the data:

Name of repository:

iDEA Repository at Drexel University

URL:

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

A.8 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, May-August, 2021. We will make necessary modification at the beginning of Phase 4, September, 2022.
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.