University at Albany, State University of New York

Data Science for the 21st Century Library and Information Professions

The Research Foundation for SUNY, University at Albany, in collaboration with the University of North Texas, seeks *$97,306* from the *Laura Bush 21st Century Librarian Program*, under the ***Lifelong Learning*** project category, and proposes a 12-month ***planning project*** that aims to identify graduate data science education models at library and information science (LIS) schools and iSchools in North America and generate preliminary findings of surveys on the extent to which perceptions, views, and attitudes of relevant stakeholders differ with respect to the skills, knowledge, values, and attitudes for an effective data science specialist and/or librarian[1]. Systematically developed and psychometrically tested survey questionnaires and focus group discussion protocol will be used to conduct the surveys. Outcomes of this project include: preliminary reports and systematically designed and psychometrically tested and refined survey instruments. The project will lay the foundation for a more comprehensive subsequent project that aims to build a model data science curriculum that will enable potential stakeholders to take advantage of the data revolution to advance their core competencies, interests, careers, and responsibilities.

1. Statement of Broad Need

For the purpose of this project, we define data science as the analysis of large datasets through computational methods and systems to generate new sources of information and knowledge (Ortiz-Repiso, Greenberg, and Calzada-Prado, 2018). We also recognize that data science cuts across multiple, but related, disciplines such as statistics, computer science, and information science. However, some consider data science a natural sub-discipline of the information sciences because none of the other disciplines make human-centered issues their main focus.

On the workforce trends and development front, over a decade ago, Hal Varian (2009) argued that in the subsequent decades, information and knowledge workers who have the skills to understand, curate, mine, visualize, and extract value from raw data by transforming it into information and knowledge as well as communicate or disseminate it effectively will be in high demand. A lack of library and information professionals with these skills, in the context of the provision of data services by libraries, has also been recognized (Corrall, Keenan, and Afzal, 2013). The data lifecycle, including processing, mining, visualization, and access, not to mention the information and knowledge derived thereof, is especially relevant to LIS schools and iSchools, their programs, and information and data professionals broadly. Therefore, placing data science programs within library and information schools is both appropriate and a strategic response to advances in technology, increased computing power, and the big data phenomenon (Marchionini, 2017). However, keeping these data science programs grounded in human-centeredness, social responsibility, and context (Shah, Anderson, Hagen, and Zhang, 2021) is paramount if LIS schools and iSchools are to build on their existing programs while adding value and options for their students.

Historically, LIS schools, iSchools, information environments such as libraries, and library and information professionals have adapted to changing landscapes with respect to new

[1] A data science specialist uses data analytics tools to mine data and applies statistical techniques and other methods such as machine learning and predictive modeling to discover relationships, trends, and insights from datasets. Data librarian is often regarded as an ad hoc term but is defined here as a librarian who provides expert support for students, faculty, and researchers in the areas of data science, data analytics, and visualization.

innovations, and the challenges and opportunities that come with it. These communities mostly acted to innovate with and integrate these opportunities into their research, teaching, and practice. The LIS field has been at the forefront of integrating technology and systems to solve practical problems in libraries and deliver services. The data revolution is one example. To respond to this revolution, data science programs have recently sprouted. However, the programs vary in their structure, emphasis, contents, and requirements. Also, there is still a data science skills gap among information professionals due to their lack of the requisite skills to work in data-rich environments where the value of a workforce with data science skills is critical for making informed decisions (Burton and Lyon, 2017). Similarly, the skills gap among data scientists who lack the traditional LIS training where topics such as data curation, preservation, metadata, etc., is another area that needs to be addressed by some current data science programs (Ortiz-Repiso, Greenberg, and Calzada-Prado, 2018). Furthermore, the effectiveness of current data science programs at LIS schools and iSchools in meeting the needs of their students and potential employers is not clear. Therefore, it is time to scrutinize these programs with respect to their ability to address and accommodate current and future workforce demands, and the interplay among employer skill demand, the acquisition of those skills, and workplace opportunities for applying the skills.

There are efforts that attempted to analyze and survey data science curricula at LIS schools and programs (e.g., Hu et al., 2017; Kang and Eunhye, 2017; Ortiz-Repiso, Greenberg, and Calzada-Prado, 2018; Si, Zhuang, Xing, and Guo, 2013; Wang, 2018; Wang and Lin, 2019) or job ads for data science positions in a single country, such as China (Wu, Lv, and Xu, 2020) or conduct bibliometric analysis of data science literature (e.g., Virkus and Garoufallou, 2019). Specifically, the work by the iSchool Data Science Curriculum Committee (iDSCC) (see Shah, Anderson, Hagen, and Zhang, 2021) parallels and complements some, if not all, of the main objectives of our project, although their project is at the planning stage.

While our project builds on these efforts, especially those that attempted to develop frameworks and models for data science education at LIS programs and schools (e.g., Song and Zhu, 2017), it is a more comprehensive undertaking with a broader scope that goes beyond analysis of information available mainly on the schools' Websites. We include in our analysis perceptions of a number of stakeholder groups on both the supply and demand sides of data science education because it is only then that we can begin to see a fuller picture. Consequently, we seek to:

1. identify data science education models through a systematic analysis of graduate level data science programs and courses offered at LIS schools and iSchools in North America; and
2. systematically develop and psychometrically test survey questionnaires and a focus group discussion schedule or protocol to conduct surveys to generate preliminary findings and reveal the extent to which perceptions, views, and attitudes of a sample of the major stakeholders (current students, graduates, and their employers or supervisors) contrast and/or complement each other with respect to the skills, knowledge, values, and attitudes that the programs were meant to instill in an effective data science specialist and/or librarian.

Through this ***exploratory*** project, we endeavor to examine and provide a current snapshot of data science in the LIS field, trends in education and workforce development, and perceptions of relevant stakeholders on both the supply and demand sides of data science education with respect to the skillsets required and expected of data science specialists and librarians. More specifically,

the main goal of the project is to assess the supply and demand sides of data science education, with the help of various methods such as surveys, content analyses, and focus group discussions to shed light on:(1) the nature of current data science programs in LIS programs and schools; (2)perceived strengths and weaknesses of graduates of the programs from the perspectives of their employers or supervisors; (3) perceptions of current students of the programs with respect to the skills, knowledge, values, and attitudes they expect to develop; and (4) gaps in the expectations of graduates of the programs and their opinions of the skills, knowledge, values, and attitudes required to be an effective data science specialist and/or librarian.

As an exploratory investigation, the outcome and findings of this project are intended to:

- produce a preliminary report on the goals/design, curriculum, course contents and structure of data science programs at LIS schools and iSchools in North America
- produce carefully designed, tested, and refined sorting and ranking activities and survey instruments together with a preliminary report summarizing the perceived skills, knowledge, values, and attitudes current data science students at LIS schools and iSchools in North America expect to develop
- produce a preliminary report of the gaps in terms of the expectations and opinions of current data science students at LIS schools and iSchools in North America with respect to the skills, knowledge, values, and attitudes needed to be an effective data science specialist and librarian/professional
- produce carefully designed, tested, and refined sorting and ranking activities and survey instruments together with a preliminary report outlining critical activities and the responsibilities, skills and knowledge required for data science specialist and/or librarianship positions
- inform and lay the foundation for a subsequent project that is broader in scope, more comprehensive, and will investigate in greater detail the extent to which the demands and needs of the supply side (LIS schools and iSchools, students, and graduates) and demand side (employers or supervisors) of data science are met; and
- identify the specific demands and needs that will eventually help us to build a model/standard graduate data science curriculum at LIS schools and iSchools that addresses those demands and needs as well as contribute to efforts that attempt to build synergy between both sides of data science education.

2. Project Design

2.1. Project Background, Goals, and Outcomes

To reiterate, the overarching goal of this project is to assess the current structure of data science programs as well as the perceptions, views, and attitudes of the supply side (LIS schools and iSchools, students, graduates) and demand side (library employers such as research libraries, special libraries, information centers, etc., and supervisors of data science specialists and librarians) of data science education to see the extent to which the two sides contrast and/or complement each other with respect to the skills, knowledge, values, and attitudes that the programs were meant to instill in an effective data science specialist and/or librarian. The four specific goals are:

1. shed light on the nature of current data science programs at LIS schools and iSchools and determine how the programs are positioned in LIS education

2. identify perceived strengths and weaknesses of the programs and graduates of those programs employed by libraries and information centers (those beyond libraries), from the point of view of their employers or supervisors
3. document perceptions of current students of those programs with respect to the skills, knowledge, values, and attitudes they expect to develop; and
4. identify gaps in the expectations of graduates of those programs who are employed by libraries and other information environments, and their current opinions with respect to the skills, knowledge, values, and attitudes required to be an effective data science specialist and/or librarian.

Previous analyses, assessments, and surveys of data science curriculum and education at LIS schools and programs (e.g., Hu et al., 2017; Kang and Eunhye, 2017; Ortiz-Repiso, Greenberg, and Calzada-Prado, 2018; Si, Zhuang, Xing, and Guo, 2013; Song and Zhu, 2017; Virkus and Garoufallou, 2019; Wang, 2018; Wang and Lin, 2019; Wu, Lv, and Xu, 2020) laid some foundation for parts of the current project. However, those are disparate efforts and vary in their scope, depth of their coverage, and analyses. All the same, our project builds on these efforts by reviewing the relevant literature to identify significant elements to create methodologically sound and systematically designed set of procedures and survey instruments that can produce valid and reliable data.

Table 1 is a summary of the goals, the methods to be used to achieve them, and the anticipated outcomes of this planning project. We anticipate that any challenges due to the ongoing COVID-19 pandemic, to the extent that the pandemic would still be a factor when the project commences in August 2021, would be minimal. This is because most of our activities, including the three questionnaire surveys, the focus group discussions, and meetings with project advisory board members can be accomplished remotely through online survey platforms such as Qualtrics and virtual meeting platforms such as Zoom. Our application for IRB approval will detail the appropriate protocol for data collection and mode of contact with human subjects, project personnel, and others that we plan to follow to comply with public health and other relevant guidelines.

Table 1: Summary of Project Goals, Methods, and Outcomes

Goal #	Methods	Outcomes
1	Content analysis of current data science courses, concentrations, and programs at LIS schools and iSchools in North America	A preliminary report on the data science programs' structures, goals/design, curriculum, and course contents
2	Survey, through a self-administered questionnaire, of a sample of supervisors or administrators at libraries (e.g., directors of the members of Association of Research Libraries/Special Library Association) and other non-library employers of data science specialists and librarians who are graduates of data science programs at LIS schools and iSchools in North America	A preliminary report outlining the critical activities and responsibilities, skills and knowledge required for data science professions and librarianship; carefully designed, tested, and refined sorting and ranking activities and survey instrument

3	Survey of a sample of current students of data science programs at LIS schools and iSchools in North America through a self-administered questionnaire	A preliminary report summarizing the perceived skills, knowledge, values, and attitudes current data science students expect to develop; carefully designed, tested, and refined sorting and ranking activities and survey instrument
4	Survey of a sample of graduates of data science programs at LIS schools and iSchools in North America, who are serving as data science specialists and librarians, through a self-administered questionnaire; focus group discussion with a sub-sample of the graduates	A preliminary report of the gaps in terms of their expectations and current opinions with respect to the skills, knowledge, values, and attitudes needed to be an effective data science specialist and/or librarian

2.2. Project Activities

This one-year project includes the following four activities (matching the above four goals).

Activity 1. Content analysis of current data science programs

To map the current state of graduate data science education, to identify key trends, and to discern areas and questions for future data science education in LIS, we will survey the websites of LIS schools and iSchools in North America. Content analysis of those schools' graduate programs, concentrations, and courses will be conducted. Academic program is defined here as any combination of courses and/or requirements leading to a degree, i.e., Bachelor's degree, Master's degree, Ph.D. degree, a certificate, or to a major, minor, or academic track, specialization, and/or concentration. Course is defined as any course offered for credit and is a required component of a student's academic program plan. However, our analysis will only be limited to graduate-level academic programs and courses because of the predominance of graduate LIS programs in North America.

With respect to the analysis of graduate data science programs at LIS schools and iSchools, first the programs will be classified based on their program type, such as degree with concentration, graduate certificate, or advanced certificate. Additionally, the scope and required credits of the program will be coded and analyzed. In our analysis of the courses in data science programs, we will review and analyze course descriptions, objectives and syllabi (when available), as well as compare course objectives, requirements, topics, assignments, projects, and other types of assessments used in the courses. Content analysis of descriptions and contents of current data science courses, concentrations, and programs at LIS schools and iSchools in North America will be conducted with the help of a coding scheme to be developed considering the concepts, topics, and issues related to data science education and practice. Intercoder consistency or reliability among two independent expert coders will be assessed through two commonly used measures, Cohen's Kappa (Cohen, 1960) and percent agreement.

While content analysis of current programs can partially reveal the nature of data science education at LIS schools and iSchools, a complete picture of the needs and demands of students, the labor market, and employers, with respect to relevant skills and competencies, can only be garnered through additional means. These include surveys of potential stakeholders. Hence the need for employing a variety of approaches including surveys and focus group discussions with students, graduates, and employers or supervisors.

Activity 2. Survey of current students of data science programs

Activity 3. Survey of supervisors or employers of data science graduates

Activity 4. Survey of and focus group discussions with graduates of data science programs

For the three surveys (Activities 2, 3, and 4), we will recruit participants and conduct surveys of the various stakeholders (students, supervisors or employers, and graduates) through various means. For instance, graduates of data science programs who are data science specialists and/or librarians will be recruited through various listservs, including lita-l, which is ALA's Library and Information Technology Association list, and sts-l, which is ACRL Science & Technology Section discussion list, to participate in the questionnaire survey and focus group discussions. The questionnaire to survey a sample of graduates will be distributed via the same platform or means, Qualtrics, as the student and employer or supervisor questionnaires. The questionnaire to survey current students of data science programs will be distributed through the deans, chairs, and program directors of LIS schools and iSchools. We will request that they forward our survey questionnaire to their current students as well as help us identify recent graduates of their data science programs.

A sub-sample of those recent graduates of data science programs who complete the survey questionnaire will be invited to participate in a focus group discussion. Focus groups are particularly suited to conduct exploratory studies and for gathering a variety of perspectives about the same topic and gaining deeper insights (Gibbs, 1997), provide an opportunity for researchers to engage deeply with participants, and probe when responses need more explanation (Barbour, 2008). Two focus group discussions (each with a group of five to eight graduates working as either data librarians or data science specialists in other non-library information environments), with the help of a carefully designed schedule or protocol, will be conducted either at conferences or using a virtual meeting platform such as Zoom to account for the challenges due to the COVID-19 pandemic. Each focus group discussion with graduates working at a variety of libraries and non-library information environments from different parts of North America will be facilitated by the PI, Co-PI, and graduate research assistant and audio-recorded, transcribed, and analyzed using a qualitative data analysis software such as NVivo.

Survey instruments such as questionnaires and the focus group interview schedule or protocol to be used to collect data on the perceptions about data science skills, knowledge, values, and attitudes by the various stakeholders on both the supply and demand sides of data science education will be developed and tested systematically. Such stakeholders include supervisors or administrators at libraries (e.g., directors of the members of Association of Research Libraries/Special Library Association), other employers of data science specialists, and current students as well as graduates of data science programs at LIS schools and iSchools in North America.

Following identification of the main constructs and variables, we will conduct an exhaustive review of the relevant literature to generate initial items and/or questions that could elicit valid and reliable responses from members of the stakeholder groups. A panel of experts, who will also serve on the project advisory board, will review the items and questions through, among other things, sorting and ranking activities, before we deploy the survey instruments. We will ensure that the survey instruments will have appropriate psychometric properties, including acceptable levels of validity (face, content, and construct validity) and reliability. Survey data will be cleaned before analyses and will be checked for correctness, completeness, duplicates, outliers, and proper formatting. Similar procedures to be employed for content analysis of

current data science programs will also be employed to conduct content analysis of data from focus group discussions with graduates of data science programs, including the creation of a coding scheme and assessment of inter-coder consistency or reliability using Cohen's Kappa (Cohen, 1960) and percent agreement.

Some of the main outcomes of this planning project are a carefully designed, tested, and refined sorting and ranking activities and survey instruments to study perceptions of the skills, knowledge, values, and attitudes of data science students, graduates, and their employers or supervisors. The goal is to produce survey instruments that will have requisite levels of face, content, and construct validity as well as reliability. While face and content validity of the instruments will be assessed both qualitatively and quantitatively (e.g., through the item impact score for face validity and the content validity ratio or CVR for content validity), their construct validity and reliability will be assessed through exploratory factor analysis and Cronbach's coefficient alpha (Cronbach, 1951), respectively.

2.3. Project Team

The project team, whose substantial experiences and backgrounds complement each other and will ensure the project's success, will consist of a Principal Investigator (PI), a Co-Principal Investigator (Co-PI), a graduate research assistant (student) to be recruited and hired, and five (5) members of a project advisory board. The PI (Dr. Abebe Rorissa) has over 30 years of experience as an LIS professional, educator, and scholar conducting several survey research projects and teaching research methods and statistics courses, supervising graduate students on their theses/dissertations, giving talks on topics related to workforce development, and leading program self-studies, evaluations, and program development efforts that introduced new, innovative, and strategic concentrations that resulted in successful accreditation as well as substantial growth in graduate enrollment. He served as Provost's Fellow where he helped create the data and institutional repository at his institution. The Co-PI (Dr. Jeonghyun (Annie) Kim), served as director of an IMLS funded digital curation and data management graduate certificate program. Her research areas include digital libraries and archives, data management and curation, and LIS workforce development. Both will collaborate to administer the project and will contribute equally toward each element of the project, including the surveys and analyses of data. The graduate research assistant (student) with expertise or research focus on data science/analytics, survey research, and qualitative and quantitative data analysis, will be recruited from the diverse pool of Ph.D. and/or Master's students of the iSchool at the University at Albany. The graduate research assistant's roles and responsibilities include assisting with most of the phases and activities of the project. The project advisory board will consist of two LIS school or iSchool directors/deans/leaders, a president/chair of a professional association and supervisor of data science specialists and/or librarians, a data science specialists and/or librarian, and a data science student at an LIS school or iSchool. Members will be recruited from a diverse pool of experts, professionals, and students in terms of their experience, leadership roles, professional networks and influence within the LIS field and information professions, and fit for the project, its activities, and outcomes. They will help review the items and questions to be included in the various survey instruments, participate in key project team meetings, provide feedback, critical review, and recommendations with respect to the development of survey instruments, data collection and analyses, assessment of the project's outcomes, and assist with planning the subsequent project that involves designing a model data science curriculum. Two members of the advisory board have provided support letters detailing their contributions to the

project and committing to collaborating with the PI and Co-PI (see Supporting Documents 2 and 3).

2.4. Project Timeline

This one-year planning project will commence on August 1, 2021 and conclude on July 31, 2022. Table 2 presents project tasks or activities to be accomplished at different points throughout the one-year period.

Table 2: Project timeline for tasks or activities

Timeline	Activities
August - October 2021 *Build instruments & get IRB approval*	Conduct a comprehensive literature review to define domains of constructs/variables/concepts in data science education and practice; Identify and define the main constructs; Generate items/questions/scales for survey instruments; Finalize formation of the project advisory board; Conduct face and content validation of survey items, scales, and/or instruments; Apply and secure IRB approval; Build survey instruments and a coding scheme to conduct content analysis of current data science courses, concentrations, and programs; Develop focus group discussion schedule or protocol; and Select samples of the various stakeholders (students, employers or supervisors, and graduates).
November 2021 - January 2022 *Finalize instruments & recruit participants*	Conduct content analysis of data science courses, concentrations, and programs; Test survey instruments; Develop refined and final survey instruments; Select stratified samples of participants of the various questionnaire surveys when sampling frames are available; and Recruit participants from the various stakeholder groups.
February - April 2022 *Collect & analyze data*	Conduct focus group discussions; Collect survey/questionnaire data from students, employers or supervisors, and graduates; Build a coding scheme to analyze focus group discussion data; Prescreen, clean, and analyze survey and focus group discussion data; Conduct assessment or evaluation of the validity and reliability of scales and survey instruments.
May - July 2022 *Finalize Project*	Write and disseminate preliminary reports based on the surveys and analyses; Finalize the remaining activities of the project; Draft a pre-proposal for a subsequent project that expands and builds on this planning project

2.5. Project Evaluation and Dissemination of Outcomes

In addition to psychometric assessments of survey instruments in terms of their face, content, and construct validity as well as reliability (including intercoder consistency), members of the project advisory board will serve as reviewers and evaluators of the design, testing, and implementation of the instruments and all the project phases and activities. Their input will be sought at critical phases of the project (e.g., survey design, sample selection, outreach to stakeholders, data collection, planning for a subsequent project, etc.) by scheduling meetings as often as needed. Results of evaluations and feedback by members of the advisory board will be used to modify and improve survey instruments and adjust project plans when possible and/or necessary.

Dissemination, to the LIS education and data science or librarian professional community as well as other potential users and stakeholders, of project outcomes such as survey data, results, and preliminary reports will be through papers and presentations at national and international conferences such as the ALISE conference, the ASIS&T Annual Meeting, the ALA Annual Conference, and the iSchool conference. Manuscripts will also be submitted to prominent journals in the field. Survey and focus group data, including the questionnaires and coding schemes, as well as any eligible pre-publication copies of manuscripts for journals and conference proceedings will be deposited into institutional repositories at the University at Albany (https://scholarsarchive.library.albany.edu/) and University of North Texas (https://library.unt.edu/scholarly-works/) for open and wider access. To the extent that it is allowed by IMLS, they will also be submitted to other platforms such as the Open Science Framework (OSF) and their locations or URLs will be distributed to discussion lists and social media sites of relevant professional associations.

3. Diversity Plan:

We will ensure that samples of survey participants are representative of the population of students, employers or supervisors, and graduates with respect to race, income, education level, gender, etc. We will ensure that members of each underrepresented group are included in the sub-sample of data science specialists and librarians who are graduates of data science programs at LIS schools and iSchools in North America and who participate in the questionnaire survey and focus group discussions. We will also consider career longevity, such as whether a participant is a new graduate or mid-career or experienced professional to include participants with diverse experiences on the job. Our sample of students will also include those who are in their first year, second year, and those who have been in the program for more than two years. Members of the board will participate in key project team meetings, provide feedback on all aspects of the project, and ensure that our diversity plan works. They will be recruited from a diverse pool of experts and group of stakeholders, professionals, and students in terms of their experience, leadership roles, position within the LIS field and information professions, and fit for the project, its activities, and outcomes.

To the extent that they are available and clearly identify demographic information, lists of members of the various stakeholders will serve as sampling frames and a representative random sample of each stakeholder group will be selected through stratified sampling where the strata match the different clusters of participants within each distinct stakeholder group. A diverse set of LIS schools and iSchools with varying sizes and from the different parts of North America, together with their data science programs, courses, and concentrations will be considered for content analyses. In the absence of diversity statistics on data science professions and librarianship, reports on diversity in the library workforce by professional associations (e.g., ALA) and statistical reports on enrollments at LIS schools and programs (e.g., ALISE) will guide our sample design that considers the diversity of students and graduates of data science programs and professionals. We also plan to engage and partner with stakeholders such as the RDAP Association, IASSIST, ALISE, and iCaucus for our sample selection and ensure that its diversity reflects the underlying population. Some members of the advisory board have provided support letters for this planning project. Cultivating the partnerships will be an integral part of the project and the board will play a very important role.

4. Broad Impact

A study of the perceptions of a particular discipline's stakeholders on the supply side of the job market such as current students with respect to programs in the discipline, their structure, usefulness, and learning environment is critical in shaping the stakeholders' approach to it (Lewis, Jackson, and Waite, 2010). Investigations into the students' skills, knowledge, values, and attitudes about their programs and discipline are also important with respect to student retention. On the other hand, misconceptions could have implications for the programs and professional growth of the graduates (Biggers, Brauer, and Yilmaz, 2008; Denning, Tedre, and Yongpradit, 2017). Ours and other similar studies could potentially allow stakeholders on both the supply and demand sides of the education sector address any misalignments and lack of synergy that exist between their needs and demands when it comes to the requisite skills, knowledge, values, and attitudes.

We envision significant impact on the nature, relevance of content or focus, and goals of future data science programs at LIS schools and iSchools in North America and beyond. Based on the baseline data, preliminary findings, and outcomes of this project, our subsequent effort will be an expanded project to conduct a more comprehensive and thorough fit and gap analysis that is broader in scope while extending and utilizing the refined and psychometrically tested survey instruments, findings, and lessons learned from the planning project as input. The ultimate goal will be to ensure that the demands and needs of both the supply side and demand side of data science education are met and to build a model or standard graduate data science curriculum with content and program goals that will:

1. serve as a basis for future data science programs that, in addition to the core contents and topics, are grounded in human-centeredness, social responsibility, and context. It will also enable the LIS schools and iSchools to think strategically when planning for growth in their programs while meeting employers' demands for professionals with the requisite skills, knowledge, attitudes, and values
2. help educators and administrators of LIS and data science programs to systematically design and improve their courses and programs to build synergy between both sides of data science education
3. prepare diverse LIS student populations for 21st century workplace roles such as data science professions and librarianship
4. equip libraries and other information environments to not only innovate their services and respond to the changing needs and demographic composition of their users through data-related services, but also supply them with a diverse cadre of data science specialists, librarians, and professionals that will help widen their scope by adding processes and services that rely on data analytics
5. give data science specialists and librarians avenues for upgrading their skills and knowledge by enrolling in systematically designed and relevant data science programs that provide them with competitive advantage, thereby providing lifelong learning and career opportunities; and
6. enable all these and other stakeholders to take advantage of the data revolution to advance their core competencies, interests, careers, and responsibilities.

Project Title: Data Science for the 21st Century Library and Information Professions; **Start Date:** 8/1/2021; **End Date:** 7/31/2022

Schedule of Completion

Activity	2021					2022						
	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul
Conduct a literature review to define domains of constructs												
Identify and define the main constructs												
Generate items for survey instruments												
Finalize formation of the project advisory board												
Conduct face and content validation of survey instruments												
Apply and secure IRB approval												
Build survey instruments & coding scheme to analyze data science programs												
Develop focus group discussion protocol												
Select samples of the various stakeholders												
Conduct content analysis of data science programs												
Test survey instruments												
Develop refined and final survey instruments												
Select samples of participants												
Recruit participants from the various stakeholder groups												
Conduct focus group discussions												
Collect survey/questionnaire data												
Build coding scheme to analyze focus group discussion data												
Prescreen, clean, & analyze survey and focus group discussion data												
Assess validity & reliability of scales and survey instruments												
Write and disseminate preliminary reports												
Finalize the project												
Draft pre-proposal for a subsequent project that builds on the planning project												

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.

The outcomes of our project are mainly survey data in the form of responses to items and questions in questionnaires and/or focus group discussions by students, professionals, and supervisors/employers. Survey data will be saved as either CSV files or other text-based formats while audio recordings from two focus group discussions will be saved as .mp4 files. The intellectual property rights of the data will lie with the two leaders of the project team, the PI and Co-PI. To the extent that it is allowed by IMLS, the dataset generated through the project will be deposited into the University at Albany's institutional repository (https://scholarsarchive.library.albany.edu/) under the creative Commons license.

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.

The intellectual property rights of the data will lie with the two leaders of the project team, the PI and Co-PI. Provided that they are allowed by IMLS guidelines, the University at Albany encourages faculty, students, and staff to make project data available publicly under the creative Commons license. Therefore, our dataset will be available for access under the following terms: users can copy, share, distribute, and adapt the dataset for non-commercial purposes. However, they need to attribute to give credit. Potential users will be notified of relevant terms or conditions through the University's institutional repository Website (https://scholarsarchive.library.albany.edu/).

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.

N/A. Data from survey questionnaires and focus group discussion notes will not contain any personal or identifiable information about participants.

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.

In addition to preliminary project reports produced in Microsoft Word and PDF formats, digital content to be collected and/or created through the various project activities include: Survey data will be saved as either CSV files or other text-based formats; audio recordings from two focus group discussions will be saved as .mp4 files.

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.

Equipment - video/audio recorder, laptops

Software - Qualtrics, Microsoft Word, Adobe Acrobat Pro, Microsoft Power Point, Microsoft Excel, SPSS, Zoom, NVivo, Express Scribe Pro Transcription Software

Supplies - External storage/disk

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.

CSV, DOCx, XLSx, mp4, PDF, txt

Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan. How will you monitor and evaluate your workflow and products?

Data collected through the questionnaire surveys will be collected and stored in Qualtrics, which can only be accessed with unique login IDs and passwords. The survey data and audio files from the focus group discussions will be downloaded and stored in a secure OneDrive folder accessed only by the PI, Co-PI, and graduate research assistant. Cleaned and processed survey and focus group data, including the questionnaires and coding schemes, will be deposited in the the University at Albany's (https://scholarsarchive.library.albany.edu/) and University of North Texas's (https://library.unt.edu/scholarly-works/) institutional repositories.

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).

Cleaned and processed survey and focus group data, including the questionnaires and coding schemes, will be deposited in the the University at Albany's (https://scholarsarchive.library.albany.edu/) and University of North Texas's (https://library.unt.edu/scholarly-works/) institutional repositories.

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).

N/A

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

N/A

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).

N/A

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).

Cleaned and processed survey and focus group data, including the questionnaires and coding schemes, will be deposited in the the University at Albany's (https://scholarsarchive.library.albany.edu/) and University of North Texas's (https://library.unt.edu/scholarly-works/) institutional repositories and made available for open 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.

University at Albany, Scholarly Archive (https://scholarsarchive.library.albany.edu/); 2. University of North Texas, Scholarly Works (https://library.unt.edu/scholarly-works/)

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.

Project data will include contents of Websites of LIS schools or iSchools with graduate data science programs and program/course descriptions, syllabi, and responses by supervisors or administrators, student, and graduates of data science programs, to items and/or questions in questionnaires and/or focus group discussion schedule or protocol. Data collection and analyses will last about six months. Results of data analyses will be used to produce preliminary reports on the extent to which perceptions, views, and attitudes of a sample of the various stakeholders on both the supply and demand sides of data science education contrast and/or complement each other with respect to the skills, knowledge, values, and attitudes that the programs were meant to instill in an effective data science specialist and/or librarian.

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 University at Albany requires all research that involves human subjects to be approved by the University's institutional review board (IRB). Therefore, in the first three months of the project, the project team will submit an application for an expedited review by the IRB and approval of the survey instruments and procedures before samples are selected and human subjects are identified.

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.

Participants of all survey questionnaires will remain anonymous, will not be required to furnish any identifying information, and will not be identified in datasets and reports. Participants of the focus group discussions will only be identified for data collection purposes and the dataset can only be accessed by the PI, Co-PI, and graduate assistant. Information collected will only be shared publicly in aggregate form with all identifying information removed. The University at Albany's institutional review board will monitor the study and may inspect the dataset.

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

The University's Qualtrics system, which can only be accessed with unique login IDs and passwords, will be used to create the questionnaires and distributed to potential participants. Qualtrics allows data to be downloaded in a CSV format that can be read and processed using any spreadsheet software. Data from focus group discussion will only include audio files to be transcribed by the project team. Questionnaire data will be analyzed using statistical software such as SPSS.

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?

The University at Albany's consent form, to be adapted to this project and approved by the IRB, will be used to inform potential participants of their rights and seek their voluntariness. The consent form will have: names of the researchers; description of the research; description of human subject involvement; risks & discomforts of participation; measures to be taken to minimize risks and discomforts; expected benefits to subjects or to others; confidentiality of records/data; contact information of the lead project team and the IRB; their rights as a research participant; voluntary nature of their participation; and withdrawal of subjects and data retention. Participants need to agree to these terms to participate in the survey.

A.6 What is your plan for managing, disseminating, and preserving data after the completion of the award-funded project?

Datasets from this project will be archived and only be disseminated through the University at Albany's (https://scholarsarchive.library.albany.edu/) and University of North Texas's (https://library.unt.edu/scholarly-works/) institutional repositories. Both Websites for the two institutional repositories have detailed policies and guidelines for managing and providing access to deposited datasets.

A.7 Identify where you will deposit the data:

Name of repository:

1. University at Albany, Scholarly Archive; 2. University of North Texas, Scholarly Works

URL:

1. https://scholarsarchive.library.albany.edu/; 2. https://library.unt.edu/scholarly-works

A.8 When and how frequently will you review this data management plan? How will the implementation be monitored?

Our data management plan will be reviewed throughout the data collection and analyses process plus at the time of deposit to the two institutional repositories and whenever the University makes changes to its policy and guidelines that will affect the integrity and availability of the dataset under the creative commons license.