ABSTRACT: Investigating the National Need for Library Based Topic Modeling Discovery Systems

The University of Notre Dame is seeking an IMLS planning grant to convene a diverse community during the 2019 calendar year that will contribute to conceptualization of digital tools that support the creation and dissemination of cross-disciplinary research. The grant would enable us to conduct a series of workshops as venues for the collaboration of domain experts, librarians, and computer science specialists. The goals are to understand their unique current practices and to identify possibilities to use topic modeling and NLP to enhance or augment current library classification in order to meet current cross-disciplinary research needs. The target community includes small and large and public and academic libraries and institutions.

The main components of the project include a comprehensive literature review, an environmental scan, and a series of workshops to begin a national conversation and to form a national team to support further research. The planning project should be considered as the foundation for further development of an innovative approach that we believe will not only support interdisciplinary scholarship, but also leverage traditional library skills and enhance digital stewardship.

The planning team intends to use these activities to:

- determine the national need for an automated tool that supports cross-disciplinary research
- identify and solidify potential partnerships for the future collaborative development of tools to support those needs
- determine how to provide a framework for librarians to engage in innovative approaches to supporting research
- engage a conversation among scholars, computer scientists, and librarians to reconcile domain-specific, best approaches to supporting cross-disciplinary discovery

The planning team will produce a white paper based on our findings and if the results from the community engagement prove positive about the need and interest, the next step will be to organize a diverse steering committee comprised of interested institutions that attend the workshops. This steering committee will be charged with developing a comprehensive plan for the next phase of the project - to conduct a research program on how to best apply what the team has learned to extend the foundational elements of *Convocate* to any cross-disciplinary research. As a part of this effort, the cross-institutional committee will apply for a research grant to design an optimized workflow for developing automated metadata and classification for cross-disciplinary discovery. The long term goal is for the workflow and research to be folded back into the platform underlying the *Convocate* project, but will require deeper research first to determine the approaches.

List of Key Project Staff and Consultants

1. **John Wang**: *Associate University Librarian for Digital Access, Resources, and Information Technology, University of Notre Dame*
2. **Christina Leblang**: *Project Manager on Catholic Social Thought and International Human Rights Law, Center for Civil and Human Rights, University of Notre Dame*
3. **Mark Dehmlow**: *Program Director of Information Technology, University of Notre Dame*
4. **Helen Hockx-Yu**: *Program Manager for Digital Product Access and Dissemination, University of Notre Dame*
5. **Eric Lease Morgan**: *Digital Initiatives Librarian in the Center for Digital Scholarship, University of Notre Dame*
6. **Alex Papson**: *Unit Head of Metadata and Digital Service and Digital Production, University of Notre Dame*
7. **Don Brower**: *Digital Library Infrastructure Lead, University of Notre Dame*
8. **Laurie McGowan:** *Digital Project Manager, University of Notre Dame*
9. **Mark Graves**: *Visiting Research Assistant Professor, University of Notre Dame*
10. **Pamela Graham**: *Director of Global Studies and Director of the Center for Human Rights Documentation and Research, Columbia University Libraries*
11. **Ed Fox**: *Professor of Computer Science and Director of the Digital Library Research Laboratory, Virginia Polytechnic Institute and State University*

University of Notre Dame - Narrative

Investigating the National Need for Library Based Topic Modeling Discovery Systems
IMLS Planning Grant Application, 2nd Round

Brief Project Description

The University of Notre Dame seeks funding for national stakeholder engagement to assess the value of machine learning and natural language processing tools to facilitate automated metadata creation and classification in support of cross-disciplinary research.

Statement of National Need

As automation provides more efficiency for traditional library functions, libraries aim to find new ways to provide value for their organizations while also fulfilling their mission to serve patrons in learning, creative inquiry, research, and knowledge/information management. The profession as a whole is transforming from supporting research by acquiring scholarly resources and making them accessible to becoming more collaboratively immersed in the creation of scholarship itself. Simultaneously, as universities increasingly focus on cross disciplinary research, libraries find that their mature discovery tools do not semantically traverse multiple, disparate academic domains. The work to support cross-disciplinary research is itself a diverse intersection of professional concerns - expertise in classification typically held by librarians, deep understanding of domain research areas that is held by scholars, and strong competency in computer learning and natural language processing typically held by computer scientists.

Convocate

As a proof of concept in this interdisciplinary vein, the Hesburgh Libraries at the University of Notre Dame partnered with the University's Center for Civil and Human Rights (CCHR) to help the institution bridge the gaps in cross-disciplinary discovery on a project called *Convocate*. The result of the *Convocate* project is a research tool for investigating the connections between international human rights and Catholic social teaching. It searches theological and legal documents simultaneously, returning results in such a way that users are able to compare these bodies of thought side by side. Additionally, the database content is parsed into granular components. Search results are returned at the paragraph level, enabling users to retrieve more contextual specific results within a document. Users also have the option of viewing the pertinent paragraphs within the full context of the document. The idea for *Convocate* arose from researchers' needs to bridge the fields of Catholic social teaching and international law. Traditionally, successful cross-disciplinary research of these two fields would require a research scholar to individually access databases of Catholic social teaching documents as well as international law documents located in a variety of international human rights organizations' websites, and then compare what they found. For example, to write coherently about child labor practices in Latin America, with reference to a strong Catholic identity and formation in social justice, a research scholar would potentially search the databases of the United Nations, the International Labour Organization, and the Organization of American States for international law documents, alongside the databases of the Vatican and the Episcopal Conference of Latin America for Catholic documents. To streamline this process, researchers wanted to create a tool that would bring documents from both fields together on one platform for simultaneous searching and side-by-side textual comparison. Scholars in the CCHR turned to the Center for Digital Scholarship,

housed in Hesburgh Libraries, and the Libraries' Web and Software Engineering team for consultation.

To a researcher, the *Convocate* research tool looks mostly like a search engine. But in actuality, *Convocate* employs multiple technologies and scholarly approaches to achieve cross-discipline discovery that include web programming and design, application and creation of domain specific vocabularies, natural language processing, machine learning, and concept mapping.

Creation of a Unique Controlled Vocabulary

The Hesburgh librarians encouraged the research scholars to use LC subject headings for document classification. After analysis of the targeted set of Catholic social teaching documents, CCHR researchers knew that most documents would be returned for almost all of the likely user created queries related to the intersection of social teaching and human rights. Thus, the decision was made to narrow in on each paragraph of text as a unique search result. Given the detailed level of searching desired, the researchers found that LC subject headings lack the specificity they wanted to apply to the texts. Instead, the researchers created a unique controlled vocabulary. They gathered a list of potential terms from common search topics on significant websites and databases, such as the United Nations and the United States Conference of Catholic Bishops. The terms from both disciplines were then merged into one coherent hierarchical list of controlled vocabulary. This list was sent to senior scholars with expertise in the intersection of international human rights and Catholic social teaching for review and critique. Finally, the controlled vocabulary was tested against the dataset by applying it for document classification. Through this last step, holes in the controlled vocabulary were identified and addressed. In addition, some terms were eliminated as they were found to be too specific and not wide found throughout the dataset.

Application of Controlled Vocabulary

Once the controlled vocabulary was fully identified, the researchers read through a set of documents from both fields applying the controlled vocabulary to each paragraph. No limits were set on the number of controlled vocabulary that could be applied to each individual paragraph. Questionable paragraphs were additionally tagged with controlled vocabulary terms. The rationale for such an approach was that other research scholars would be given the opportunity to see these paragraphs on the periphery of a particular concept and make their own decisions as to whether or not to include these texts in their personal work. Eventually, given the number of paragraphs (over 11,000) and controlled vocabulary terms (over 250), it became evident that the task of reviewing each paragraph for each term was too time-consuming. With the collaboration of librarians and computer scientists, they turned to topic modeling to semi-automate the tagging process. The controlled vocabulary terms were mapped to various topics from the topic modeling algorithms in order to tag the remaining paragraphs. Due to the complexity of the chosen hierarchical controlled vocabulary, the terms were not necessarily mapped one-to-one with the topics from the computer algorithms. Currently the team is working to create a more automated approach of tagging that will utilize the current dataset as a training module for classification of new documents as they are added to the database.

Lessons from Convocate

Both librarians and CCHR research scholars gained valuable insights through collaborative work on the *Convocate* project. These insights reinforce what many librarians view as the limitations of current library practices and they highlight the importance of controlled vocabularies, domain-subject expertise, and the use of topic modeling and other automated computer techniques for document classification.

1. Importance of Controlled Vocabulary

While both fields speak about the same concepts, these concepts are represented by different vocabulary or technical jargon. For example, in regards to labor issues, Catholic documents largely use the language of "wage," and legal documents largely use the language of "remuneration." Even so, both wage and remuneration can be found in both types of documents. To capture more fully the results that speak about compensation for work, a person would need to search both "wage" and "remuneration." Moreover, supposing that someone is more heavily immersed in the legal field, that person might only search remuneration. Search results would still be returned in both fields but the scholar would miss a significant number of results, especially from the Catholic documents. Because scholars from different disciplines speak largely different "languages" to describe similar concepts or ideas, a set of controlled vocabulary as compared to a keyword search is crucial for a successful cross-disciplinary search.[1]

Gross (2014) cited a 2010 study by Nowick, Travnicek, Eskridge, and Stein, that investigated the use of controlled vocabulary and keywords identified by automated text analysis or word clustering techniques for documents in an online environment. The study explores similarity among terms from users, from the documents themselves, and from controlled vocabularies. Their findings showed that "the controlled vocabulary terms were better matched to both users' search terms and document terms than documents to users. Correlations between users and controlled vocabularies were 2–3 times higher [than] between users and documents. This suggests that, through controlled vocabularies, libraries do provide a bridge between users and relevant documents. These results would indicate that human catalogers are the ideal way to organize documents into a library. However, given the limitations of humans to undertake a complete catalog of the internet, there may be ways to refine cluster-based organizing algorithms for digital libraries."

2. Importance of Domain-Subject Expertise

The emphasis that academic libraries place on professional cataloging competence leads them, naturally enough, to trust and consult trained and skilled cataloguers to produce

[1] Some librarians have suggested letting go of controlled vocabulary and solely relying on keyword searches. Although OCLC (2009, 2011) and some librarians, Gross (2014) and Romanowski (2016), indicated that users favor keyword searching in search engines, keywords alone cannot close the terminology gap across fields and may leave specific resources invisible to users. Under the current technological context and disciplinary "silos," controlled vocabulary remains relevant as it continues to offer value in authority weighting. New relational instruments that connect disciplines at the conceptual level are critical to adapt to the changing needs in research and scholarship. Some have created thesauri to systematically define concepts and relationships, but Qin and Paling (2005) have argued that they are less expressive and flexible compared to machine-created ontologies.

reliable and useful records. However, cataloging expertise is not the same as subject expertise, and no cataloguers can be expected to have equal depth of knowledge in every subject (Campbell & Fast. 2004). Interdisciplinarity requires key knowledge of the concepts in more than one field, as well as familiarity with theoretical methodologies from different disciplines. In response, libraries should involve scholars of those fields to establish an ontology for exploring interdisciplinary topics (Denda, 2005). For example, from the *Convocate* project, "solidarity" is a pervasive concept in Catholic social teaching that recognizes the responsibility to work for the common good of others because we are all persons. "Cooperation" is a similar concept in legal documents that addresses the need for nations to work together to solve international problems of economic, social, cultural or humanitarian nature. In creating the controlled vocabulary for *Convocate*, the researchers chose to combine these two concepts: "solidarity/cooperation." By choosing "solidarity/cooperation" from the topic list, users will be able to explore meaningful search results that reflect the common threads from the two fields about working together to improve human lives. The topic search enables a user from one discipline to overcome the problem of nuanced vocabulary in the other discipline and, hence, uncover relevant information that might otherwise remain hidden within the context of current classification schema.

3. Importance of Machine Automation/Topic Modeling

During the first phase of *Convocate's* development, the team discovered one of the most significant limitations was the arduous and time-consuming task of human creation and association of metadata. While it always will be necessary to train topic modeling systems to develop classification and concept associations, the current level of human intervention cannot scale to millions of documents. It is clear that automation will be required. Gleaning from existing scholarly research on automated metadata creation and topic modeling, we know that libraries and organizations like OCLC have made progress in this area (Golub, 2006). Caragea et al. (2014) describe an automatic classification method applied to documents harvested from the Web. Danilevsky et al. (2014) outline a framework for topical keyphrase generation and ranking, based on the output of a topic modeling of short documents. The approach of Bijalwan et al. (2014) is similar to what the team considering, namely, "first categorize the documents using KNN based machine learning approach and then return the most relevant documents."(2014) From the team's perspective, none of the literature posits methods for comparing & contrasting seemingly disparate corpora.

Ontological methodologies and semantic linking are two hot topics for better search and discovery in general and there is a bright future for strategies to aid cross-disciplinary research. Once concepts and sub-concepts are mapped and their relationships declared, computers are able to direct users to the information they seek. Ontological work requires substantial knowledge and expertise and needs more information professionals, including librarians, to engage in its development for computers and algorithms to accurately automate classification.

University of Notre Dame - Narrative

Project Design

The future of how librarians and libraries contribute to research and discovery is rife with emerging opportunities. As libraries consider innovative approaches to research tools and as cross-disciplinary research itself gains importance in the academy, now is the time to engage domain experts, librarians, and computer scientists to explore new possibilities in advancing cross-disciplinary knowledge creation based on user-centric design principles.

The team's approach with *Convocate* has proven that libraries can help combat the inadequacy of library and information science classification tools and practices in advancing multidisciplinary research. The team believes that the platform underneath the tool, as well as the natural language processing techniques they have used, could be extended to solve this problem for many different cross-disciplinary research projects. The purpose of the IMLS Planning grant would be to convene a diverse community that engages in facilitating the creation and dissemination of cross-disciplinary research. The grant would enable us to conduct a series of workshops as opportunities to invite domain experts, librarians, and computer science specialists to understand their unique current practices and explore new possibilities to advance cross-disciplinary knowledge creation based on user-centric design principles. While the initial work on *Convocate* demonstrated the value and need for cross-disciplinary approaches to discovery tools, the team feels it can have a more significant impact to extend *Convocate*'s approach to other cross-disciplines, e.g., psychology and economics, chemistry and art, etc. As shown in the suggested workshop agenda (see appendix), the participants would engage in conversations around case studies and brainstorm possible future work. The team's goals are: 1.) to determine the national need for an automated tool that supports cross-disciplinary research and 2.) to identify and solidify potential partnerships for the future collaborative development of tools to support those needs.

The team firmly believes that librarians, as experts in classification systems, taxonomy, ontologies, and semantic/linked data, are critical in influencing the evolution of classification and metadata creation. However, libraries cannot do it alone. It will be essential to involve the domain experts of various disciplines, to best design systems that can support scalability, extensibility, and sustainability. The ability of libraries to innovate knowledge management will no doubt be a determining factor in the profession's relevance to our patrons. Technological advancement has altered users' information seeking behaviors; and the changing nature of research has been calling for bibliographic classification systems' accurate and immediate response to all possibilities (Beghtol, 1998).

This planning project should be regarded as the foundation for further development of an innovative approach that the team believes will strengthen the support of interdisciplinary scholarship by advancing library knowledge classification practices into the era of machine learning and artificial intelligence. For purpose of illustration, a Notre Dame professor has already provided us with a challenging use case to be used to stimulate discussion within the workshops.

University of Notre Dame - Narrative

The proposed project plan will include the following:

1. Literature Review and Environmental Scan
2. Workshops
3. Report on Activities and Future Directions
4. Developing Formal Partnerships and Research Grant Proposal

1. Literature Review and Environmental Scan

The team has already completed initial research into the needs for enhancement to cross-disciplinary classification and automated metadata for discovery. The team has reviewed materials from the natural language processing, digital humanities, library, and other communities to determine some of the institutions actively engaged in this type of inquiry. The first part of the planning activities will be to complete the team's research and environmental scans as a means for targeting participants who can share their work and ideas in the workshops. This particular activity will be one of two activities to fuel participant ideas and enrich the workshop discussions and outcomes (see Appendix for workshop outline).

2. Workshops

Dates: 4 dates in Spring 2019

Locations:

- New York City, NY (Columbia University),
- South Bend IN (University of Notre Dame),
- Washington D.C.,
- Palo Alto, CA

Target Participation: 20 - 25 people from a diverse background of academic and public libraries, large and small universities, HBCUs. Participants will be determined through a combination of open call and invitation. Invitees will be determined through library, computer science and domain scholar professional networks and experts identified in the literature review.

Theme: *Automatic Classification of Documents & Cross-disciplinary Research*

Evaluation: Formative evaluation survey after each - output will inform revision to next workshop; final workshop output to be included in summative evaluation.

<u>Goals:</u>

- Explore national need for automated classification systems in support of cross-disciplinary research
- Find a diverse group of potential partners to engage in further research and development.

- Build a diverse national community of shared interest to ensure the team understands disparate community needs and structure so that the team's subsequent work can accommodate differences in institutional size, focus, and mission.

3. Report on Activities

Once all of the three information gathering activities are complete, the team will compile a white paper on the need for, and interest in, a national community to support automated metadata classification for cross-disciplinary discovery. The paper will be shared with participants in the workshops and broadly through the internet using various channels:

- library technology discussion lists
- digital humanities online communities

The research and report will also be synthesized into an article for publication in a scholarly journal.

4. Developing Formal Partnerships and Research Grant Proposal

If the results from the community engagement prove positive, the next step will be to organize a diverse steering committee comprised of interested institutions that attend the workshops. This steering committee will be charged with developing a comprehensive plan for the next phase of the project - to conduct a research program on how to best apply what the team has learned to extend the foundational elements of *Convocate* to any cross-disciplinary research. As a part of this effort, the cross-institutional committee will apply for a research grant to design an optimized workflow for developing automated metadata and classification for cross-disciplinary discovery. The long-term goal is for the workflow and research to be folded back into the platform underlying the *Convocate* project, but will require deeper research first to determine the approaches. With robust and diverse community input into the needs for and recommended developmental directions for automated cross-disciplinary classification, the subsequent research project will better mirror articulated community needs across a broad constituency of institutions and therefore have broader applicability for the cultural heritage community.

Other Cross-Disciplinary Applications

The interdisciplinary methodology and user interface framework of *Convocate* can be applied to other interdisciplinary research endeavors. Jeffrey Bergstrand, a finance professor in the Mendoza College of Business at the University of Notre Dame and a Research Associate of CESifo, an international network of researchers based in Europe, has worked with several colleagues to assemble a database of economic integration agreements from 1950-2012. His

research spans issues of economic liberalization, intellectual property rights, welfare, governance etc. Notably, there are strong connections between economic agreements of countries and their different governance structures. Thus, a potential suitable fit within the proposed methodology would be the dataset of economic integration agreements as compared with constitutions and other government documents. Overcoming semantic challenges between the larger disciplines of economics and political science would provide a more coherent picture of nation states. Utilizing these two larger datasets, researchers could explore issues of globalization answering questions about how governance structures impact international trade agreements and vice versa. Jeffrey Bergstrand's work also touches upon welfare and standards of living as understood from an economics point of view. He has suggested another cross-disciplinary dataset could be economic resources and those in psychology with a focus on answering questions about subjective well-being and life satisfaction as compared to economic well-being. Other cross-disciplinary datasets that have been suggested from research scholars include questions surrounding medical ethics as understood from a medical, philosophical, and theological perspective and questions involving indigenous aboriginal communities as answered from western historical accounts as compared to aboriginal accounts.

Diversity

With this planning grant, and later software development phases, the team is taking an intentionally inclusive approach. The workshop locations were chosen for a broad geographic dispersion - West Coast, Midwest, and East Coast. The locations also are easy to get to and within driving distance of a substantial and diverse collection of academic and public libraries. As part of the environmental scan and initial outreach to the community, the team performing additional targeted outreach to both small and large and public and academic libraries. The team's experience is that larger libraries often heed the call to be involved in exploratory and open source efforts, and the team wants to ensure to capture the voice, and potential partnership, of libraries of all sizes to ensure that the final result can have the broadest applicability. To date, the team has elicited confirmed interest from a Center for Black Literature and Culture that is part of the Indianapolis Public Library as well as a historically black college. The team has collegial connections to small colleges and other underrepresented academic institutions and will actively solicit their participation to include multiple perspectives on the need for development of innovative systems that will leverage machine learning for the support of cross-disciplinary research.

The team feels there is an unarticulated gap, particularly in open source software development in libraries, where smaller institutions are seldom involved, due to financial and staffing constraints. The team wants to identify ways in which long term goals can be formulated to include these considerations so that smaller institutions can participate in the future development of a broader application, and development can account for the differences among communities

University of Notre Dame - Narrative

National Impact

While *Convocate* serves as an example of how library organizations can contribute to cross-disciplinary research, the team's long-term ambitions are to extend the method, model, and functionality so that it can be used for any cross-disciplinary research. The team believes this work can have a significant impact both for scholars and libraries if it is extended to additional cross-disciplinary fields. For researchers, the team believes that scaling the capabilities of the platform and broadening its scope of cross-disciplinary content to science, social sciences, and humanities, could augment library science theory, knowledge, and practices with computational methodologies for supporting discovery and scholarship. Moreover, for libraries, there is a significant opportunity to leverage traditional library skills such as classification, metadata creation, and disciplinary stewardship in ways that create new value and augment skills development of of library employees such as catalogers, subject liaisons, and digitization specialists. to help transition into the knowledge work of the future. The team believes, additionally, that this work provides an opportunity to revitalize the perception of libraries as collaborators in the overall scholarly endeavor.

Since the planning grant intends to determine the national need, success should be measured by several dimensions:

1. The project team will be able to demonstrate diversity in the representative user groups that it plans to invite. Participants will be made up of scholars, computer scientists, and librarians from academic institutions of varying size and resources.
2. The team will ensure that participants are chosen to represent a variety of use cases and practices in the series of workshops.
3. The workshops will be structured to lead participants to raise a rich set of questions and potential solutions.
4. The planning grant will identify a likely cohort and partnerships that will help library professionals pursue further collaborative effort in improving the cross-disciplinary research agenda.

One of the deliverables is to document current thoughts, models, practices, and tools. This will contribute to the professional literature and hopefully inspire more peers to jump-start similar projects at their institutions.

Summary Statement

The future of how librarians and libraries contribute to research and discovery is rife with emerging opportunities. Given that libraries are expanding the ways in which they can support research and the depth of their expertise in classification, ontologies, and metadata, the team feels that taking applying those skills to innovative approaches can transform the ways in which research tools can support our end users. The team will need to work closely with a diverse set of experts from different areas of expertise and different types and sizes of organizations if the team wishes to have the greatest possible impact. The team believes that the first step in this effort is to build that diverse community and to engage in discussions that can determine how to optimally move forward in this endeavor. The team feels that an IMLS planning grant is just the tool to help us create the community and discourse to accomplish that.

SCHEDULE OF COMPLETION

	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov
Activity 1: Literature Review and Environmental Scan												
Activity 2: Workshops												
Activity 3: Report on Activities and Future Directions												
Activity 4: Develop Formal Partnerships and Research Grant Proposal												

DIGITAL PRODUCT FORM

Introduction

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

Instructions

D Please check here if you have reviewed Parts I, II, III, and IV below and you have determined that your proposal does NOT involve the creation of digital products (i.e., digital content, resources, assets, software, or datasets). You must still submit this Digital Product Form with your proposal even if you check this box, because this Digital Product Form is a Required Document.

If you ARE creating digital products, you must provide answers to the questions in Part I. In addition, you must also complete at least one of the subsequent sections. If you intend to create or collect digital content, resources, or assets, complete Part II. If you intend to develop software, complete Part III. If you intend to create a dataset, complete Part IV.

Part I: Intellectual Property Rights and Permissions

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

We will create a website that is openly accessible, and will continue to share and publish meeting minutes, literature, and a workshop recordings and documents via an Open Science Framework (OSF) project. All resources will be openly licensed for re-use and redistribution . We plan to publish in open journals, and/or to secure open access right of the final paper or the right of depositing a preprint copy to our intentional repository with commercial publishers. All workshop participants will be informed about the above IP guidelines and will be expected to exercise their right to openly share any direct or derivative works of this federally funded project.

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.

My institution commits to publish all digital content open access. All workshop participants will be informed about the above IP guidelines (outlined in A.1) and will be expected to exercise their right to openly share any direct or derivative works of this federally funded project.

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

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

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

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

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

We plan to use Google Docs to record meetings; leverage the Zoom application to record lectures and lightening talks. Participants may choose Microsoft Office or Google word-processing and presentation software to produce their workshop materials. We have chosen OSF to store and/or link to any project documents. Anyone who has access to a web browser will be able to access to all content of the project.

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

Recordings will be in MPEG-4 format, captured via Zoom. The software allows a single-resolution with fixed frame rate format to ensure better compatibility with various media players.
Meeting minutes will be captured in Google Doc Format
Lectures will be produced in Microsoft docx, Macintosh Page, or Google doc format
Presentations will be created in Microsoft ppt, Machitosh Keynote key, or Google Slides format

B. Workflow and Asset Maintenance/Preservation

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

N/A

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

We will leverage OSF for public access to the project documents, lecture recordings, and meeting minutes. OSF is available for free to the public and allow for versioned updates. For preservation of project documentation, there could be risks of format obsolescence. We will convert documents to PDF/A or txt format; choose to preserve a legacy media player to play recordings or update the video formats to the future standards. Files can also be deposited into our institution repository.

C. Metadata

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

OSF uses Dublin Core and extensions to Dublin core. Although metadata is arguably not subject to copyright in the United States, to facilitate the use metadata across borders and avoid later arguments or confusion, we will make metadata available under a CC0 license. A simple policy statement effectively noting that to the extent that there is copyrightable expression, that copyright interest is being made fully available through a CC0 grant. Such a statement would acknowledge that there may not be any copyright protection in the metadata while simultaneously ensuring that any doubt is removed over whether such metadata may be shared or reused. It removes legal risk for those who are seeking to use the metadata, particularly because of the existence of a large number of institutional policies relying on a CC0 license for metadata. OSF uses and supports thesauri based on:

- Taxonomy of Research Doctoral Programs from the National Academies
- Classification of Instructional Programs, 2010 edition, from the National Center for Educational Statistics (NCES)
- Medical Subject Headings (MeSH) from the National Library of Medicine
- Law subject headings are informed by the Current Index to Legal Periodicals (CILP) and FindLaw
- Several Business categories come from Cabell's

Metadata can be customized with discipline specific terms if wished by participants. When we deposit files into our institution repository, we will provide any necessary, technical, descriptive, administrative, and preservation metadata for a single item or an archival package.

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

We are using OSF for preserving data and metadata and this information will be maintained via the University of Notre Dame during and after the award period in our repository. We can transfer assets between OSF and CurateND using packages compliant with the bagit specification.

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).

Both OSF and our repository expose data to search engines. We will ensure that our sites are open for indexing by search engine bots, so any materials can be easily discoverable by the general public.

D. Access and Use

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

The digital content will be publicly available via OSF and eventually be preserved in the repository. Users only need a web browser, a plugin for media players and document viewers. For example, a user may open multiple word-processing documents, such as docx, pdf, rtf, and stream videos of multiple formats, such as mp4, mov, and wmv.

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

https://convocate.nd.edu/
https://curate.nd.edu/

Part III. Projects Developing Software

A. General Information

N/A

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.

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

B. Technical Information

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

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

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

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

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

C. Access and Use

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

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

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

Name of publicly accessible source code repository:

URL:

Part IV: Projects Creating Datasets

N/A

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

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?

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

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

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

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

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

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

Name of repository:

URL:

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