DADAlytics: A Tool to Steer Digital Culture Heritage to the Semantic Web

Prepared by M. Cristina Pattuelli, Pratt Institute, School of Information

Abstract

The School of Information at the Pratt Institute is seeking a National Leadership Grant for Libraries, National Digital Platform, Planning Grant with the goal of devising a sound methodology, including testing and preliminary prototyping, for DADAlytics — a completely re-engineered version of the Linked Jazz Transcript Analyzer, an open source tool that performs computational analysis on textual documents to generate linked open data and knowledge graphs. The project will be led by Professors Cristina Pattuelli and Matt Miller with linked data specialist Karen Hwang and will bring together stakeholders from Carnegie Hall Archives, Tulane University Digital Library, University of Minnesota's Umbra Search African American History, the Whitney Museum of American Art's Research Resource Department, and the Harvard University Center for Italian Renaissance Studies, Villa I Tatti. Partner institutions will provide insight and input on desired functionality, including processing workflows and user needs. They will also contribute samples from their collections for testing and to ensure the tool works in a variety of professional settings and is able to process different types of archival and special collection materials.

Beginning October 1, 2017 and running through September 30, 2018, the project activities will include: **(1)** an environmental scan that surveys current methods and tools for computational text analysis and identity management; **(2)** a full-day meeting with stakeholders, the outcome of which will be a prioritized list of system requirements and staff and end user needs to inform the design of DADAlytics; **(3)** iterative testing and prototyping sprints culminating in the development of a proof of concept or preliminary prototype; **(4)** documentation and dissemination of project results; and **(5)** the completion of a work plan for further development and implementation of DADAlytics.

This project responds to the need for intuitive semantic tools that help expose digital cultural heritage content held in discrete institutional repositories. One of the most promising semantic technologies, linked open data, makes it possible to identify meaningful relationships between documents beyond institutional boundaries and across domains, giving cultural heritage greater visibility and discoverability. To date, the methods and tools used to generate linked data require significant technical expertise. We envision DADAlytics to be an open source linked data service that significantly lowers the barrier to generating and publishing high quality networked data for a broad range of cultural institutions and information professionals, as well as humanities scholars and researchers. With input from institutional partners and key stakeholders, we believe such a service has the potential to become an integral part of the collection processing workflow of every digital project regardless of scope, size, platform or budget. Every librarian, archivist, museum professional and digital humanities scholar has the potential to contribute to the linked open data environment, offering unprecedented opportunities to advance scholarship and create new knowledge.

Narrative

1. Statement of National Need

Troves of digital cultural heritage resources are held in libraries and archives often hidden from potential user communities. Recent reports have stressed the urgency for cultural institutions to make their archival and special collections—often their most distinctive and valuable resources—more accessible to the public. These same reports have also highlighted the challenges associated with meeting rising user expectations in terms of access and use. As a method for publishing and interlinking data from diverse and heterogeneous sources, linked open data holds enormous potential to address the problems posed by data currently siloed in databases and institutional repositories, one of the most urgent challenges in the development of a national digital platform (IMLS Focus, 2015)[1]. However, decreasing budgets and the limitations of current documentation practices have prevented all but the most well-resourced cultural heritage institutions from introducing the technical expertise needed to incorporate linked data into their workflow settings and practices and significantly enhance users' access to collections.

Text exploration and mining through machine learning and natural language processing (NLP) are powerful techniques for harnessing structured and unstructured content and data to discover hidden and new knowledge, affording users unprecedented opportunities for new forms of engagement with and analysis of content. Textual analysis performed with automatic methods is a well-established practice in Digital Humanities research. Named entity recognition (NER), for example, is frequently used to support digital history and literary research. Automatic approaches to building relation networks from text have been applied in numerous projects, from newspaper archives (Ardanuy et al., 2016)[2] to works of fiction (Elson, Dames, and McKeown, 2010).[3]

New approaches that apply computational methods for the creation of additional value to archival collections have been introduced, including Thomas Padilla's The Carl Woese Collection[4] work, TOME (Interactive TOpic Model and MEtadata Visualization)[5], Ed Summers' Fondz[6], and ArchExtract[7] at UC Berkeley's Bancroft Library. Unfortunately, as Ellings (2016)[8] points out, these tools have failed to achieve wide adoption. With a focus on born-digital archival materials,

[1] IMLS Focus. (2015). *IMLS Focus summary report: National Digital Platform for libraries, archives, and museums.* Retrieved from https://www.imls.gov/sites/default/files/publications/documents/2015imlsfocusndpreport.pdf

[2] Ardanuy, M. C., Knauth, J., Beliankou, A., van den Bos, M., & Sporleder, C. (2016). Person-centric mining of historical newspaper collections. In *International Conference on Theory and Practice of Digital Libraries* (pp. 320-331). Springer International Publishing.

[3] Elson, D. K., Dames, N., & McKeown, K. R. (2010). Extracting social and intellectual networks from literary fiction. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 138-147). Association for Computational Linguistics).

[4] Padilla, T. (2013). Topic Modeling for Digital Collections Appraisal. *Society of American Archivists Annual Meeting*. New Orleans, Louisiana.

[5] http://dhlab.lmc.gatech.edu/tome/

[6] https://github.com/edsu/fondz/blob/master/setup.py

[7] https://github.com/j9recurses/archextract

[8] Ellings, M. (May 24, 2016). Using NLP to support dynamic arrangement, description, and discovery of born digital collections: The ArchExtract Experiment [Blog post]. Retrieved from https://saaers.wordpress.com/2016/05/24/using-nlp-to-support-dynamic-arrangement-description-and-discovery-of-born-digital-collections-the-archextract-experiment/

the ongoing project BitCurator NLP[9], based at the University of North Carolina at Chapel Hill, is developing NLP software to create reports on relevant features (e.g., preservation, information organization and access activities) found in text through entity extraction. Also centered on born-digital resources, the project ePADD[10], based at Stanford University's Special Collections & University Archives, now in phase 2, has developed a software tool to support email archiving, processing, and discovery. ePADD applies NER methods to extract named entities in the archives (person, organization or locations) mentioned in email messages and reconciles them with name authorities.

Initiatives are also emerging that combine automatic computational methods and linked open data technologies using the actual content of digital resources as the source of linked data. A notable example is provided by the project Pelagios Common[11] that employs NER methods for acquisition and identification of geographic names in non-English historical texts. In the library community, most of the work on linked data development has focused on the conversion of legacy data—such as MARC records and EAD finding aids—into linked data to enhance resource discovery. As models for resource descriptions evolve, interest is growing in finding new ways to leverage record data, including the use of automatic text analysis for text mining of bibliographic records (Godby, Wang and Mixter, 2015)[12]. In the field of archives, one notable project is SNAC (Social Networks and Archival Context),[13] which employs NER to leverage the standardized structure of EAD finding aids to generate EAC-CPF (Encoded Archival Context-Corporate Bodies, Persons, and Families) records in order to expose the social connections between people and historical records.

Automatic text analysis methods and linked data technologies are at the core of the Linked Jazz Project.[14] Under the direction of Pattuelli and Miller, project leads for this proposal, Linked Jazz is an ongoing project based at the Pratt Institute that has pioneered successful methods and tools for generating and visualizing linked data derived from digitized oral history transcripts from various jazz history archives. With generous support from OCLC, New Orleans Jazz and Cultural Heritage Foundation, and Ella Fitzgerald Charitable Foundation-JEN Research Fellowship at the Smithsonian, the Linked Jazz team has used the funding to help researchers and educators uncover the social relationships between musicians, offering new paths of discovery and interpretation of primary source materials. A key component of the Linked Jazz set of tools is the Transcript Analyzer[15]. Developed by Matt Miller in 2012, this tool performs named entity recognition and identity management for linked data generation and enables us to process interview transcripts and create a curated open dataset of social networks that has been reused by other music-related projects, such as JazzCats (Jazz Collection of Aggregated Triples).[16] Over the years, we have received numerous requests to adapt the Transcript Analyzer for other types of digital cultural heritage projects from such institutions as Berklee College of Music, The Harvard Center for Italian Renaissance Studies, and The Pina Bausch Foundation, as well as individual scholars and

[9] https://www.bitcurator.net/bitcurator-nlp/

[10] https://library.stanford.edu/projects/epadd

[11] http://commons.pelagios.org/about/

[12] Godby, J., Wang S. and Mixter, J. K. (2015). Library Linked Data in the Cloud: OCLC's Experiments with New Models of Resource Description. *Synthesis Lectures on the Semantic Web: Theory and Technology*, *5*(2), 1-154.

[13] socialarchive.iath.virginia.edu/index.html

[14] https://linkedjazz.org/

[15] https://linkedjazz.org/tools/transcript-analyzer/

[16] http://jazzcats.oerc.ox.ac.uk

music librarians such as Steve Ellingson, professor of Sociology at Hamilton College, and Nick Paterson, music librarian at Columbia University.

In general, methods and tools to generate linked data pose technical and workflow challenges to librarians and information professionals, especially in institutions with limited budgets. The need for in-house expertise and the lack of user-friendly tools prevent the cultural heritage community from fully participating in the production cycle of linked data and contributing their digital collections to the linked data ecosystem. In order for linked data to scale in the area of cultural heritage, we believe that more intuitive tools, that are designed to fit seamlessly into the workflow of virtually any digital project, are needed now more than ever. The process of transforming relevant bits of information found in structured or unstructured text into high quality linked data to be published and interlinked beyond institutional boundaries and across domains must be made as simple and straightforward as possible. Capitalizing on the lessons learned and building on the team's expertise and experience in linked data development for libraries and archives, this IMLS grant would enable us to better understand how the Transcript Analyzer could be re–engineered to become a low-barrier linked data service capable of processing a broader range of text-based documents, while remaining flexible enough to work across different platforms, domains and usage contexts.

Finally, generating and connecting linked data from text holds enormous potential to open new avenues for creative exploration of cultural heritage and for unanticipated lines of research inquiry. We have only begun to envision the new research questions, methods of analysis, and creative scholarship that can be elicited when we are able to provide integrated access to cultural heritage data and metadata. As such, exploring ways linked open data can become part of the digitization process of every cultural heritage institution directly aligns this project with the National Digital Platform (NDP) funding priorities to advance the digital capacities of libraries and archives nationwide. Representing and interconnecting the information contained in digital cultural heritage content fully embodies significant NDP priorities, including "increasing access to digital services, expanding the range and types of digital content available, improving discoverability." At the heart of the linked open data initiative lie open Web standards and the capacity to generate openly available data that will facilitate sharing and reuse. This tenet further aligns this project with the NDP's goal of facilitating interoperability, both syntactic and semantic, and "supporting open access," contributing to a culture of open scholarship and "radical and systematic collaborations" between librarians, technologists and researchers.

2. Project Design

The aim of this planning project is to enable our development team to rethink the Linked Jazz Transcript Analyzer's existing software architecture and, with critical input from key stakeholders, lay the groundwork for a complete system redesign that will enable DADAlytics to support new workflows and a broader range of material types. We currently envision DADAlytics to be composed of a series of customizable modules combined to create a linked data service that allows new modes of data generation, access and dissemination through local, state, regional and national platforms, such as DPLA and Umbra Search African American History[17]. DADAlytics will be web-based and open source to maximize ease of use and adoption.

[17] https://www.umbrasearch.org/

This project capitalizes on the strong competencies of the project leads and team—their record of innovation as well as their ability to bring highly technical yet easy-to-use work to fruition. It will begin with a review of current methods, NLP software libraries and tools for computational text analysis and identity management to identify suitable components to be used in the development of DADAlytics. To this end, a preliminary survey of the current NLP landscape has been conducted and is available in Supporting Document n. 1 – Appendix A.

With the goal of maximizing the usability and flexibility of the service, a diverse pool of stakeholders will be involved in a full-day planning meeting. Subsequent individual meetings, in-person and virtually, will also take place at critical stages of the project development to share results and gain input on testing and prototyping sprints. The purpose of the initial meeting is to identify and assess technical and user requirement specifications including desired functionality and processing workflow to inform the design of DADAlytics. Stakeholders were approached for their support on the basis of their expertise, background and existing partnership. They represent a variety of cultural institutions and will bring different perspectives on professional practices, collection processing workflows, and end user needs. In addition, stakeholders will contribute sample documents to be used in the project testbed. More information on project participants is available in the section 2.2 Personnel and Participants.

A significant percentage of time and effort will be devoted to analysis and testing activities with the goal of identifying the optimal software components and usability features DADAlytics will be built upon. Within the framework of current evaluation research, including the work of Van Hooland et al. (2015)[18], Gangemi (2013)[19], and Rodriquez, Blanke, and Luszczynska (2012)[20], we will test open source NLP software libraries and computational analysis tools (from well-known Stanford CoreNLP[21] to more recent TensorFlow Syntaxnet[22]) on various corpora and document types to assess technical capabilities, performance and usability. Features considered for testing include document formats and APIs supported, amount of expected pre-processing and supervision, ability to work as customizable modular components, required infrastructure and support for web hosted solutions.

DADAlytics is intended to perform two key functions: text analysis and identity management in order to generate RDF triples through a workflow from text to knowledge graphs. The planning activities will help identify optimal technical and usability features as well as suitable software components to be integrated into the final tool to mine text for named entities and create meaningful links between these entities. DADAlytics would apply NER methods to digitized text for locating and extracting relevant named entities and then assign Uniform Resource Identifiers (URIs) to the individual tokens. Based on the type of co-occurrences of the entities detected in

[18] Van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T., & Van de Walle, R. (2015). Exploring entity recognition and disambiguation for cultural heritage collections. *Digital Scholarship in the Humanities, 30*(2), 262-279.

[19] Gangemi A. (2013). A comparison of knowledge extraction tools for the semantic web. In: Cimiano P., Corcho O., Presutti V., Hollink L., Rudolph S. (eds) The *Semantic Web: Semantics and Big Data. ESWC 2013*. Lecture Notes in Computer Science, 7882, pp. 351-366. Springer, Berlin, Heidelberg.

[20] Rodriquez, K. J., Bryant, M., Blanke, T., and Luszczynska, M. (2012). Comparison of named entity recognition tools for raw OCR text. In *Proceedings of KONVENS 2012*, Vienna, pp. 410–414.

[21] https://nlp.stanford.edu/

[22] https://github.com/tensorflow/models/tree/master/syntaxnet

text, RDF triples will be generated that represent relations among those entities. To automatically identify relevant entity associations, the inherent structure of a document is to be leveraged. A distinguishing trait of the Transcript Analyzer, precursor to DADAlytics, was its capability to break down the content of documents into relevant structural units. For example, when working with interview transcripts, pairs of questions and answers were recognized and assigned to either an interviewer or an interviewee. This served as the basis for identifying personal connections, the focus of Linked Jazz, as recorded in the text. One of the goals of this project is to augment the capabilities of the Transcript Analyzer to process a broader range of the types of documents typically found in digital archives and special collections, from transcribed oral history to letters, from diaries and personal narratives to theater booking ledgers. To this end, the involvement of stakeholders will be critical in this planning phase to gain their perspective on types of documents to process, range of entities and relationships to uncover, and features to implement.

Identity management is another essential task DADAlytics will need to support. Name resolution is at the core of linked data practices and addresses the problem of identifying whether two resources refer to the same real world entity. Multiple are the sources of variations of names, from homonymy and synonymy to spelling mistakes and OCR-induced errors, common in digitized texts. In the linked data context, resolving and disambiguating named entities within or across documents requires associating named entities located in text to a corresponding URI from a name authority (e.g., VIAF, ULAN, etc.) or a knowledge base (e.g., DBpedia, Wikidata, etc.) in the linked data cloud. For the Transcript Analyzer, we addressed the task of disambiguating homonyms, detecting inaccuracies, and assigning standard identifiers through an embedded tool[23] mapping to the Library of Congress Name Authority File (NAF) and VIAF (see Supporting Document n. 2 - Appendix B.1). To maximize the quality of the data output, we combined the automated approach with human supervision consisting of manually validating matches when multiple options occurred (see Appendix B.2). To facilitate manual supervision we prototyped an external service, called Ecco!,[24] an Italian term that emphasizes quick and effortless delivery (see Appendix B.3). While a few other identity management tools for linked data already existed, Ecco! was rather unique in that it offered web-based access and an intuitive user interface that gave librarians and other users the ability to contribute to the data curation process in a collaborative, distributed, and incremental way.

We will adopt a similar approach with DADAlytics in order to make a tedious and labor-intensive task like name reconciliation easy to perform, ensuring that results are immediately sharable and reusable. Based on input from stakeholders, we will also consider including support for minting new identifiers when no URI is available and thus contributing to generating local authorities. This feature has the potential to be of great value for projects dealing with local names (e.g. less-known people or organizations) as it makes it possible for this data to be properly published and reused. In the past few years, a number of NLP tasks have been configured for supporting entity resolution and several entity extraction APIs do now provide services for named entity extraction and disambiguation for linked data development and applications for major linked data hubs. More name vocabularies have been released in linked open data format covering various areas of interest (e.g., ISNI, Wikidata, EAC-CPF and the Getty vocabularies). Lessons learned from extensive piloting conducted with Ecco!, along with input from stakeholders and testing of available tools

[23] https://linkedjazz.org/tools/name-mapping-tool-and-curator/

[24] https://linkedjazz.org/ecco-a-new-tool-for-collaborative-named-entity-resolution/

will help identify desirable features and requirement specifications to include in our linked data service to support projects of different scale and nature.

2.1 Deliverables

The planning project will have the following deliverables:

- A proof of concept or preliminary prototype. All development byproducts including source code, scripts and sample data resulting from testing and prototyping will be made publicly and freely available through the cloud repository service GitHub for reuse, collaboration and modification.
- Consolidation of partnerships with stakeholders to lay the ground for future collaborations on the next iteration of project development.
- Working documents including survey, meeting and research notes, finding reports, model workflows disseminated through a public blog where we will also share progress updates and other content on topics of interest.
- Proposal draft delineating work plans for further DADAlytics development in anticipation of pursuing future grant funding.
- Dissemination of project results through public presentations, scholarly publications as well as blog posts and social media throughout the grant period and beyond. All publicly stored research outputs generated using this grant will be stored under an MIT or Creative Commons license.

2.2 Personnel and Participants

The project includes **M. Cristina Pattuelli** (Principal Investigator), **Matt Miller** (Technical Lead) and **Karen Hwang** (Data Specialist). **Pattuelli** will be responsible for project oversight, supervision of project staff, and relationships with stakeholders. She will provide her expertise with methodology and project design and will be in charge of project documentation and dissemination. **Miller** will provide technical expertise to conduct analysis and implementation of system requirement specifications and perform iterative testing and prototyping**. Hwang** will conduct the environmental scan, create the project testbed, and perform data analysis and evaluation in coordination with Miller and Pattuelli. A full-time **Graduate Assistant** from Pratt School of Information will join the team for the duration of the project to coordinate project activities including the environmental scan, organization of the planning meeting, communication with external liaisons, online communication (project updates via blog and social media outlets) and data archiving.

In addition to the project team members, we have identified a core group of stakeholder representatives from diverse cultural institutions that will participate in the planning meeting, contribute document samples and provide input and feedback on testing and prototyping outcomes, as discussed earlier in this proposal. These partner organizations have been chosen because of the shared interest in advancing access and use of their collections through the lens of linked data and mutual interest in collaborating.

In the context of the Linked Jazz Project we have an ongoing partnership with Carnegie Hall Archives[25] which hold the historical collections of this iconic performance venue in New York

[25] https://www.carnegiehall.org/History/Carnegie-Hall-Archives/

City including thousands of concert programs, posters and fliers, musical manuscripts and autographs going back to 1891 and are now in the process of publishing their performance history database as linked open data. Their involvement with this planning project stems from their keen interest in applying NER methods to two major collections: booking ledgers which have been recently digitized and about to be publicly released through DPLA and a series of letters of musicians and biographical portraits of relevant value for music history (see Supporting Document n. 3 – Letters of Support).

Tulane University Digital Library[26] in New Orleans is another existing partner with extensive collections documenting the history of the city. In recent years we have processed several transcribed oral histories from their Hogan Jazz Archive through the Linked Jazz Transcript Analyzer. During the process, issues of local names with no existing URIs available have posed interesting challenges. This would be one of areas that we have mutual interest in addressing through this planning project (see Supporting Document n. 3 – Letters of Support).

Umbra Search African American History,[27] based at the University of Minnesota, provides access to domain specific collections aggregated across the country with significant contribution of DPLA materials. An informal collaboration between Umbra and Linked Jazz teams has been going on from the very beginning of this initiative. The main focus has been on modeling aspects, including identifying appropriate predicates to represent connections between entities and documents from the two projects' datasets. As an aggregator service with a variety of staff users as well as communities of end users, Umbra offers a unique scenario where issues related to platform integration including workflow settings and practices can be analyzed and tested with their input (see Supporting Document n. 3 – Letters of Support).

The Harvard University Center for Italian Renaissance Studies, Villa I Tatti[28] in Florence holds the archive of Bernard Berenson, the Italian Renaissance art historian and critic and of his wife Mary Berenson, also an art historian who worked in the shadow of her more renowned husband. Mary Berenson Papers include a rich collection of letters, personal diaries, literary journals and notes, both published and unpublished, that have yet to be fully explored. In line with their ultimate goal to uncover and semantically represent the vast network of friendships and intellectual relations surrounding the couple, the Center will contribute samples from their collection of Mary's thirty diaries, spanning over fifty years (see Supporting Document n. 3 – Letters of Support).

The Whitney Museum of American Art[29] in New York City will be another key participant in this planning project. Building on a continued collaboration through student fellowships established between the two organizations and conversations with colleagues at the Frances Mulhall Achilles Library and Archives on new strategies to enhance access to their impressive collections of archival and documentation materials, this organization will be engaged with contributing document samples as well as input and feedback throughout the testing and prototyping cycle. Because they are in the process of acquiring high profile art archival collections, including primary

[26] https://digitallibrary.tulane.edu/
[27] https://www.umbrasearch.org/
[28] http://itatti.harvard.edu/
[29] http://whitney.org

sources very diverse as for typology and document structure, they will be in the position to offer particularly valuable use cases from the domain of artist archives (see Supporting Document n. 3 – Letters of Support).

As follows is a list of participant stakeholders. Noted with an asterisk are those who will receive travel funding under this grant.

Ilaria Della Monica, Archivist – Villa I Tatti, The Harvard University Center for Italian Renaissance Studies, Florence, Italy

Rob Hudson, Manager Archives – Carnegie Hall, New York, NY

Cecily Marcus* - Principal Investigator, Umbra Search African American History, University of Minnesota, Minneapolis, MN

Jeff Rubin* – Digital Initiatives and Publications Coordinator, Tulane University, New Orleans, LA

Farris Wahbeh – Benjamin and Irma Weiss Director of Research Resources, Whitney Museum of American Art, New York, NY

2.3 Work Plan

We are requesting $50,000 in grant funds to support work for 12 months from October 1, 2017 through September 30, 2018. The project will be implemented in four phases, beginning with reviewing computational analysis methods, tools and services, proceeding with a full-day meeting with key stakeholders, followed by testing and prototyping activities and ending with documentation and dissemination of results. A timeline for the phases is outlined in the Schedule of Completion.

Phase 1: Set Up and Environmental Scan (October 2017 – November 2017)
We will begin by conducting background investigation on system and user requirements, including an environmental scan reviewing NLP software libraries, services, and tools suitable for the development of DADAlytics. Karen Hwang will be in charge of this phase in collaboration with the team and with support from a graduate assistant. A draft of results and open questions, including a cognitive walkthrough of the Linked Jazz Transcript Analyzer, will be prepared for the planning meeting.

Phase 2: Defining Requirements and Planning Meeting (November 2017 – December 2017)
Project team will convene a full-day Planning Meeting drawn from the core participants listed in section 2.2. The meeting will take place at the Pratt Institute's Manhattan Campus (New York, New York) during fall 2017. Participants include representatives from Carnegie Hall Archives, Tulane University Digital Library, University of Minnesota's Umbra Search African American History, the Whitney Museum of American Art's Research Resource Department and Harvard University Center for Italian Renaissance Studies' Villa I Tatti. Travel funding will be available for two participants from outside New York. The participant from Villa I Tatti, Italy, will join remotely. Target outcome includes working documents to inform the next project phase.

Phase 3: Testing, Prototyping, and Assessment (January 2018 – July 2018)

The focal element of this phase will be testing and prototyping activities informed by stakeholders input and feedback throughout the project. The technical development will be performed by Matt Miller, responsible for the creation of the Linked Jazz set of tools in team with linked data specialist Karen Hwang. This phase is also concerned with preliminary results analysis and expected further testing and refinements with a special focus on the data processing workflow, platform integration and usability features. The entire team will be involved in analysis and outcome assessment.

Phase 4: Documentation and Dissemination of Results (June 2018 – September 2018)
The final months will be devoted to finalizing the analysis of the project results and to producing documentation including drafting a proposal outlining future implementation plans and strategies. Dissemination of results, including use cases and lessons learned, will be shared through conference presentations, scholarly publications as well as blog and social media outlets, extending beyond the grant period.

3. National Impact

There is a growing consensus around the very tangible benefits linked open data technologies could have if systematically applied to the volume of digitized cultural heritage materials available today. The involvement of librarians and archivists in linked data production, however, is key to the scalability of any linked open data project, yet is still limited to those organizations with the budget and staffing required for innovation. The willingness of librarians, archivists and museum professionals 'to get their hands dirty' and take full advantage of semantic technologies has been hampered by the lack of easy-to-use tools that fit into their day-to-day collection processing workflows.

The impetus behind this project is the desire to develop a tool that would lower existing technical barriers and make linked open data creation and publication simple and something that is seamlessly integrated into existing documentation practices. Information professionals and digital humanities researchers with only the most basic familiarity with the principles and techniques of linked data development could be at the forefront of the emerging web of data initiative moving "library and archival services in the US forward," an imperative of the DNP.

The innovative approach we propose — leveraging computations methods to generate linked data directly from content — holds enormous potential to expand access to and use of the digital content and unique collections in the U.S., another DNP priority. It is not difficult to envision the impact that exposing cultural heritage materials to new, more granular analysis would have on our historical record. Uncovering new relationships and discovering meaningful patterns, as well as relating information across sources and beyond institutional silos will greatly enhance and expand our understanding of our nation's history and culture.

Currently, automatic computational methods are not easy to use, especially by end users. Engaging stakeholders in the early stage of development is key to the development of tools and services that can and will be used. As part of the planning project, the pool of information specialists we have brought together from a variety of cultural institutions and communities of practice will increase our understanding of the needs, practices and challenges of our intended community of staff and end users. Our preliminary investigation on existing computational methods and tools, in combination with feedback on system and user requirements from stakeholders is intended, first and foremost, to inform the preliminary stages of development of DADAlytics, ensuring a sound methodological foundation.

The development activities of testing and piloting will result in a proof of concept or preliminary prototype. Additional outcomes from this early stage of development will include test cases, training data, scripts and source code, all made available open source and shared on GitHub to facilitate sharing and reuse. All the products of this grant will be published under an MIT or Creative Commons license.

The knowledge acquired through the planning process is expected to be rich, including usability, methodological and technical facets. Broadly shared through documentation and dissemination via multiple channels, as detailed in the work plan, this body of work is expected to be of significant value for our professional community. It will further our understanding of the benefits and challenges of automatic computational techniques when applied to digital cultural heritage. It will also contribute to the professional literature on semantic innovation for memory institutions as well as to establish best practices in this area of development.

Schedule of Completion

	2017			2018								
	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
PHASE ONE Set up and environmental scan	▨	▨										
PHASE TWO Defining requirements and planning meeting		▨	▨									
PHASE THREE Testing, prototyping, and assessment				▨	▨	▨	▨	▨	▨	▨		
PHASE FOUR Documentation and dissemination of results									▨	▨	▨	▨

DIGITAL PRODUCT FORM

Introduction

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

Instructions

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

PART I: Intellectual Property Rights and Permissions

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

Pratt Institute will be the copyright holder of software and sets of data produced as part of the project to ensure the application of a non-restrictive license for the material. Software developed as part of the project will be released under the MIT license due to its simplicity and lower boundaries to reuse. Datasets will be released under a Creative Commons (CC) license, which is more applicable to datasets and can be tailored to the type of contributions coming from partnering institutions. For example, to distribute a dataset resulting from a tool evaluation that also includes the source material we could use whichever type of CC license the partner is most comfortable with.

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 organization will assert ownership of the products in order to apply non-restrictive licenses and facilitate dissemination and sharing. No additional terms or conditions would be applied beyond the provided licenses to gain access to the materials.

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

Project results, including tool evaluation outcomes, are likely to be accompanied by the source materials contributed by partnering institutions. We will work with each institution to ensure that proper permissions have been secured before making content public.

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.

N/A

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.

N/A

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).

N/A

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).

N/A

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).

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

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).

N/A

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.

N/A

Part III. Projects Developing Software

A. General Information

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

The software developed will consist of scripts used for testing and rapid prototyping processes performed on existing tools, using document samples and text corpora provided by partner stakeholders. The scripts will utilize various libraries to perform operations required to support the future development of the DADAlytics data service.

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.

The software created as part of this planning project will be used to implement existing open source software libraries, including Stanford NLP and NLTK for testing and prototyping with the goal to incorporate suitable components into DADAlytics. While some customization might be needed to fit the requirements of different types of documents as for their structure and domain, overall the software developed will not create new functions.

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.

The software will be written in Python 3.0. Python was chosen for its compatibility with the software libraries we have started to consider for this project (see Appendix B: Preliminary Survey). Many of these libraries are written in Python, while those in other languages, like Java, have native bindings available for Python.

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

We will perform various rounds of software implementation to assess the suitability and effectiveness of the software for our testing and prototyping activities.

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

Development will be done on Linux and OSX systems. However any system that supports Python and the software libraries that will be used in our development activities should be able to use the software developed.

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

The software will be developed in sprints based on the software libraries and tools we will work with and the customization required to apply them to our testbed. Scripts will be documented at the source level as well at the tool level with README files that explain how to execute the implementations for the various libraries employed.

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

A good example of these types of implementation scripts is offered by our Linked Jazz Project. See the tools released for processing datasets at https://github.com/linkedjazz/linked-jazz-name-directory. All of our tools have been made publicly available throughout their development process via https://github.com/linkedjazz.

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.

No ownership rights or conditions for access and use will be imposed by Pratt Institute. The MIT software license has been chosen for this project as it is one of the least restrictive licenses for data reuse and will make any software developed accessible to the largest audience possible. All data outcomes, including scripts and documentation materials, will contain the MIT license text to clearly communicate the conditions under which the software is available.

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

The software developed will be mostly used for evaluation purposes. It will be made available for others to duplicate our testing and prototyping processes as well as for any other use the public might find useful, such as the demo of an implementation for a specific NLP library. We will host the software developed on GitHub in public repositories for ease of use and open access by the public. However, GitHub is a private company. While we recognize its great value for community collaboration and engagement, it might not be suitable for long term hosting. To this end, we will also make the deliverables available on Pratt Institute web servers as well as available on an archive.org (Internet Archive) repository.

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

For the duration of the project we will use https://github.com/linkedjazz to deposit all source code. At the conclusion of the project, the code will be available via Pratt Institute's servers (https://www.pratt.edu) and The Internet Archive (https://www.archive.org).

Part IV: Projects Creating Datasets

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

We will collect document samples from our partner institutions that will be for testing and prototyping. The creation of the testbed will occur in fall 2017, after the planning meeting with partner stakeholders. Sets of data resulting from development activities will be produced at different stages of the project (for details see Work Plan) and published on GitHub as they become available.

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?

No IRB is required.

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

No PII data will be collected.

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.

Any additional documentation related to use or restrictions of use of the sampling documents and testbed will be part of the publication datasets including appropriate licensing.

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).

Data will be generated as the outcome of testing and prototyping activities performed on existing NLP libraries and computational analysis tools. Any script used to generate data will be released so that the entire data production process could be investigated and/or replicated. The data schema generated to support data modeling and generation, a plain text JSON or CSV data file, will be also shared and documented. In addition, the methods applied to create the project testbed, including document sampling collected from partners, will be described in the documentation.

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?

Results from testing and prototyping, including data schemas and modeling workflows, will be documented as part of the project deliverables. Documentation of research results will be shared primarily through a public blog and through README files on GitHub. At the conclusion of the project, the public blog will continue to be maintained and remain live, while also saved as web captures on The Internet Archive (https://www.archive.org) for long-term availability.

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

The datasets as well as the software produced will be hosted on GitHub, Pratt Institute servers and Internet Archive for long-term availability.

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

For the duration of the project datasets will be available at https://github.com/linkedjazz. At the conclusion of the project, datasets will also be available via Pratt Institute's servers (https://www.pratt.edu) and The Internet Archive (https://www.archive.org).

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

The data management plan will be systematically reviewed throughout the duration of the project. The project team will ensure that data archiving will be completed by the end of the project cycle and as part of the project deliverables.