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

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

Summary. The School of Information at Pratt Institute (Pratt SI) is seeking a National Leadership Grant for Libraries, National Digital Platform, Planning Grant with the goal of developing DADAlytics, an enhanced version of the Linked Jazz Transcript Analyzer, a modular tool that performs supervised entity extraction from archival documents for generating linked open datasets. The funds will support the research and data gathering needed to inform the redesign and reengineering of the tool, including an environmental scan, a series of meetings with key stakeholders and the development of a prototype.

Background and Project Rationale. This project builds upon the expertise and the experience our team has gained through the development of Linked Jazz, an ongoing project 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. While most libraries, archives and museums (LAMs) engaged in the linked data initiative focus on converting bibliographic or archival metadata into linked data format, Linked Jazz leverages the intellectual content of the textual resources themselves to develop deeper and more efficient ways of processing digital cultural content. The precursor to DADAlytics is the Transcript Analyzer, developed in 2012 for the Linked Jazz Project. This tool has enabled us to process over fifty interview transcripts and to create a curated open dataset of social networks that has been re-used by other music-related projects such as JazzCats. Over the years we have received numerous requests to adapt the Transcript Analyzer for other LAM and digital humanities projects from such institutions as Berkelee College of Music, The Harvard Center for Italian Renaissance Studies, The Pina Bausch Foundation, and the Umbra Search African American History at the University of Minnesota. Capitalizing on the lessons learned and the extensive feedback we have received, this IMLS grant would enable us to better understand how the Transcript Analyzer could be re–engineered to become a data service capable of processing a broader range of text-based documents, while remaining flexible enough to work across different domains and usage contexts. In order for LOD to scale in the area of cultural heritage, we need intuitive tools that information professionals can use to convert relevant bits of information found in text, such as people, places, events and facts, into high quality linked data. We aim to make DADAlytics an intuitive service that any librarian, archivist, museum professional, or digital humanist can use regardless of their level of technological fluency. We envision DADAlytics to be composed of customizable modular services that form a complete tool. Yet these modules could be integrated into data ingest pipelines and enrichment services for local, statewide or national platforms, including DPLA. DADAlytics will be web-based and open source to maximize access and adoption.

Potential Impact. 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 available to the public while pointing out the challenges of meeting rising user expectations in terms of access and use. These

collections, both digitized and born digital, are at the center of several statewide and national initiatives to provide broad and integrated access—DPLA first and foremost among them. Linked Open Data in libraries, archives and museums is just beginning to have an impact on the way cultural institutions enhance access to their collections to enable new kinds of research, discovery and analysis. This has been recognized as one of the most urgent challenges in the development of a national digital platform (IMLS Focus Summary Report: NDP, 2015). We aim to make DADAlytics a service that significantly lowers the barrier to linked data creation. If designed well with input from institutional partners as well as practitioners, such a service has the potential to become an integral part of the knowledge organization workflow of every digital project.

Project Goals and Outcomes. A planning grant will allow the project team to:

1. Conduct an environmental scan to gather background information. The scan will help assess comparable applications and tools and inform decisions on redesign and development.
2. Leverage existing partnerships with Carnegie Hall Archives, The Hogan Jazz Archive at Tulane University and Umbra Search at University of Minnesota to identify and assess potential user needs and desired functionality and design requirements for DADAlytics.
3. Develop a prototype and conduct testing. The technical development will be performed by Matt Miller, responsible for the creation of the Linked Jazz set of tools.
4. Document and disseminate project results through scholarly publications and social media.
5. Forge additional partnerships (e.g., Smithsonian Jazz Oral History Program, The Harvard Center for Italian Renaissance Studies) and draft a proposal for a future implementation grant.

Proposed Work Plan.
The project will run from October 1, 2017 to September 30, 2018:

Phase 1	Oct-Dec 2017	Environmental scan and initial meeting with partners.
Phase 2	Jan-June 2018	Development of a prototype based on findings from Phase 1. A second meeting in June 2018 with stakeholders will be held to share the prototype and conduct preliminary testing with the goal of outlining implementation plans and strategies.
Phase 3	June-Sept 2018	Documentation and dissemination of results. Begin drafting a proposal for a full implementation grant of DADAlytics.

Personnel. M. Cristina Pattuelli, Associate Professor at Pratt SI and founder and director of Linked Jazz, will lead the DADAlytics project. Matt Miller, technical director of Linked Jazz and Head of Semantic Applications and Data Research at the New York Public Library, will provide technical expertise and develop the prototype. Pattuelli and Miller will work with a team made up of two experienced linked data specialists, Karen Hwang and Hannah Sistrunk, project manager Rachel Egan, and research assistants from the Pratt SI User Experience Team.

Estimated Budget. The project partners request $48,600 to support the project goals listed above, covering the costs of an environmental scan, planning meetings, and prototype development and testing.