(1) dLOC as Data: A Thematic Approach to Caribbean Newspapers

(2) List of team members, titles, and roles on the project

Project Leads
Miguel Asencio, Executive Director, Digital Library of the Caribbean (dLOC), Florida International University; Senior administrative lead.
Jamie Rogers, Assistant Director of Digital Collections, Florida International University; Senior administrative lead.
Perry Collins, Scholarly Communications Librarian, University of Florida; Project lead
Hadassah St. Hubert, CLIR Postdoctoral Fellow in Data Curation for Latin American and Caribbean Studies, Florida International University; Scholarly lead

Partners

Florida International University Libraries (FIU)
The following participants will focus on OCR evaluation, geo-location, named entity extraction, text analysis and dataset preparation, and developing a preservation solution.
Rebecca Bakker, Digital Collections Librarian
Molly Castro, Digital Humanities Librarian
Boyuan Guan, Lead Developer, GIS Center
Jill Krefft, Institutional Repository Coordinator

University of Florida Libraries (UF)
The following participants will focus on preparing newspaper data for distribution and analysis.
Chelsea Dinsmore, Director, Digital Support Services
Laura Perry, Digital Production Manager, Digital Support Services
Laurie Taylor, Chair, Digital Partnerships & Strategies
Caribbean Data Curation Graduate Intern (see Appendix B for position description)

Advisory Committee
The following participants will advise on corpus development and dissemination, data documentation, and outreach activities such as local community training events and edit-a-thons.
Julio Capo Jr., Associate Professor of History, Florida International University
Fletcher Durant, Head of Conservation and Preservation, University of Florida
Alex Gil, Digital Humanities Librarian, Columbia University
Melissa Jerome, Project Coordinator for the Florida & Puerto Rico Digital Newspaper Project, University of Florida
Amalia Levi, Archivist and Cultural Heritage Professional, HeritEdge Connection in Barbados
Preeya Mohan, Fellow, Sir Arthur Lewis Institute of Social and Economic Studies, University of the West Indies, St. Augustine
Leah Rosenberg, Professor of English, University of Florida
(3) Investigator Bios

Miguel Asencio is the Director of the Digital Library of the Caribbean (dLOC). He has trained teams of technicians and staff on digitization projects, ensured that quality control and environmental standards for production and facilities were met. Asencio is a digitization specialist well-versed in preservation and archiving standards in the United States (FADGI) and in Europe (Metamorfoze). His advanced degrees, both completed and in progress, have enabled him to create instructional material for use online. His interest in Curriculum and Instruction: Learning Technologies led him to develop and implement K-12 and post-secondary outreach programs, which include developing thematic collections designed to increase the use of the digital library for teaching and research. He has led numerous digitization, preservation and archiving trainings in the United States and abroad. In his capacity as Director of the Digital Library of the Caribbean (dLOC), Asencio has worked closely with partners in the United States and across the Caribbean to foster productive and beneficial collaborative relationships. In this role he assesses partner needs and finds solutions to issues that are often unique to organizations operating in the Caribbean and Latin America.

Jamie Rogers is the Assistant Director of Digital Collections at Florida International University (FIU). In this capacity, she leads the digital production, digital scholarship, data management strategies, and preservation for internally and externally funded digital initiatives in collaboration with the FIU community, as well as local partners, including municipalities, cultural institutions, government agencies, and scientific organizations. She has curated and managed over 100 digitized special collections and institutional repository collections, which are accessed an average of 9 million times per year. Since 2009, she has served as PI and Co-PI for thirteen successful grant and local partner initiatives, including projects sponsored by the Institute of Museum and Library Services and the Society of American Archivists, amounting to over $1.5 million in funding. She holds a M.S. in Management of Information Systems from Florida International University.

Perry Collins is the Scholarly Communications Librarian at the University of Florida in Gainesville, where she manages initiatives promoting open access in scholarship and education, copyright literacy, ethical approaches to digital scholarship, and capacity building for born-digital library publishing. Before joining UF in 2018, Collins held a similar position at the Ball State University Libraries in Muncie, Indiana, and worked for six years as a program officer in the Office of Digital Humanities at the National Endowment for the Humanities. While at the NEH, Collins played a major role in administering the grant review process and shaping funding programs at the intersection of technology and the humanities. She also co-managed the NEH-Mellon Humanities Open Book Program, an effort to digitize out-of-print scholarly monographs and disseminate them under open licenses. Collins holds a M.L.I.S. from the University of Illinois at Urbana-Champaign and M.A. in American Studies from the University of Kansas.

Hadassah St. Hubert, Ph.D. is currently the CLIR Postdoctoral Fellow in Data Curation for Latin American and Caribbean Studies with the Digital Library of the Caribbean (dLOC) at Florida
International University. She received a Ph.D. in History from the University of Miami and her dissertation, *Visions of a Modern Nation: Haiti at the World’s Fairs*, focuses on Haiti’s participation in World’s Fairs and Expositions in the twentieth century. Hadassah served as the Assistant Editor for *Haiti: An Island Luminous*, a tri-lingual digital humanities site dedicated entirely to Haitian history and Haitian studies. *An Island Luminous* pairs books, manuscripts, newspapers, and photos digitized by libraries and archives in Haiti and the United States with commentary by more than 100 authors at 75 universities around the world. As a Postdoctoral Fellow with dLOC, she leads programming and digitization efforts in collaboration with dLOC’s partner, L’Institut de Sauvegarde du Patrimoine National (ISPAN). In this cooperative project, she has provided training and expert technical assistance to ISPAN in its digitization efforts. In addition, she has secured over $500,000 in funding for dLOC partners for various digitization projects.

(4) Summary of Project
Digital Library of the Caribbean (dLOC) intends to enhance access to its existing Caribbean newspaper collections by making texts available for bulk download to its users. This will facilitate modes of scholarship that depend on access to image and textual data at scale and will enable a new level of access to titles not included in newspaper data resources such as Chronicling America. To meet the needs of the dLOC community for teaching and research, we will demonstrate the potential of newspaper data by creating a pilot thematic tool kit focused on hurricanes and tropical cyclones. The toolkit will provide multilingual datasets focused on these disasters from several countries and islands in the Caribbean, such as the Bahamas, Belize, Cuba, the Dominican Republic, Grenada, Haiti, Jamaica, and Martinique. The dataset collection from these newspapers have coverage from different periods of time and can provide scholars with insights into Caribbean culture and society as well as the role of resiliency within disasters.

(5) Project Rationale and Statement of Significance
*Rationale & User Communities*
Digital Library of the Caribbean (dLOC) is a multi-institutional, international digital library that has worked on data curation and digitization projects with archives and libraries across the Caribbean. dLOC recognizes Caribbean institutions’ ownership of their cultural/national patrimony, while providing access to scholars and students around the world. Scholars, practitioners, and students engage with dLOC not only as an access point for digital objects, but also as a shared node that supports public scholarship and pedagogy. As dLOC collections have grown to include almost 4 million pages and over 75 institutional partners, there is an immediate need to facilitate computational analysis to enable new modes of storytelling and collaboration.

Administered by Florida International University (FIU) in partnership with the University of the Virgin Islands (UVI), dLOC’s online technical infrastructure is currently provided by the University of Florida (UF). The Caribbean Newspapers subcollection makes up about 25 percent of the total number of pages in dLOC, with titles published in 21 countries in nine distinct languages, dating from 1783 to 2019. dLOC has the largest digital collection of Caribbean Newspapers available
online in a single platform. UF already contributes to Chronicling America and has received NEH funding to make available newspapers from Puerto Rico and the Virgin Islands. In addition, several Endangered Archives Programme grants to dLOC partners have made dozens of newspapers from various countries in the Caribbean available to users. We propose an initiative that will complement these projects and focus on Caribbean newspapers as a broad data source that lends itself to a thematic approach.

The project will undertake three related goals:

1. Enhance access to existing, previously digitized Caribbean newspaper collections by making titles available for bulk download and by better documenting available methods for collecting batch files. The project will focus on curating Caribbean newspaper data as a crucial source of political, social, and economic history across the region, with opportunities to home in on the lived experiences of individuals and communities over time. (See Appendix A for a draft list of titles for this project.) dLOC’s Caribbean Newspapers site already acts as a robust counterpart to resources such as Chronicling America and Europeana Newspapers. However, it does not offer the same level of access for those pursuing computational analysis of newspaper page images or textual data. These files are currently available externally for dLOC newspapers only via web scraping at the page level; this project will refine a workflow for extracting, packaging, and documenting assets to enable simpler access for users with a range of technical expertise. Through FIU’s open access data repository, the collections data will be cross-searchable, discoverable, and harvested to multiple access points including DataCite and Google/Google Data, supporting a broad audience.

2. Develop a pilot thematic toolkit that showcases the potential for computational analysis of newspapers as a lens onto the history of hurricanes and tropical cyclones and their impact on the region. This toolkit will include a relevant subset of the underlying textual data; one or more structured datasets derived from the original data; and a descriptive document or finding aid with information about data assets and examples of how they might be used. In this case, we plan to extract portions of text referencing hurricanes and tropical storms. Building upon this foundational dataset, we will experiment with both named entity recognition and manual techniques in order to develop a linked data model that establishes relationships between identifiable storms and specific people and locations. While we envision a series of thematic toolkits to be developed in the future, during the grant period we will specifically focus on coverage of hurricanes in Caribbean newspapers.

3. Finally, we will emphasize local capacity building and community engagement throughout the project and beyond the grant period through our faculty and teacher trainings. In the context of this project, “local” includes both development of infrastructure at the lead applicant institutions as well as partner nodes across the dLOC member community that are
creating, contributing, and reusing materials but may lack adequate support. This will include development of a long-term outreach strategy around these particular data resources and the impact of opening up dLOC data to new kinds of analysis. We will emphasize training opportunities both for our core project team and for others in the dLOC community by supporting local community events.

These project aims were conceived with the needs of dLOC’s distinct user communities in mind. A relatively small community of dLOC users has both the subject expertise and technical knowledge or programming expertise to process available data and undertake large-scale text or image analysis. These users are more comfortable starting with less structured data from a variety of sources to conduct computationally intensive analyses, and they can more easily identify external data sources to augment and enhance dLOC collections (e.g. historical hurricane data alongside historical newspaper reports). This community is most likely to seek access to page images or text files at scale as an initial starting point for research.

Based on the project team’s experience, a much larger community of scholars, educators, and collecting institutions are eager for a lower-barrier model that would package and interpret structured data as a starting point for scholarly or classroom use. These users are comfortable experimenting with plug-and-play software (e.g. Timeline JS, Palladio, Neatline) and have a working knowledge of how to identify appropriate datasets for such tools. The needs of this community drive our decision to develop thematic toolkits that offer curated, derived datasets that can be used with open-source software without any specialized technical knowledge. The toolkit’s interpretive framework and training modules will promote understanding of how the dataset was produced as well as potential gaps or pitfalls in the data.

Existing Resources & Needs

Data sources: dLOC’s collection of about 1 million digitized and born-digital newspaper pages offers rich coverage of topics across geographic, linguistic, and temporal boundaries. Because many newspapers in the collection were digitized ten or more years ago, OCR quality varies significantly, though more recent digitization efforts have greatly benefited from UF’s adoption of ABBYY Finereader for OCR processing. The project will rely on language experts on the advisory committee to document the quality of current OCR resulting from papers in French, Spanish, English, and Dutch (including some that include more than one language), as well as potentially Papiamento and Haitian Kreyòl. This project will help prioritize which newspaper titles are most in need of reprocessing. Our work with this multilingual corpora and documentation of OCR quality will be useful for other institutions with collections from non-English speaking nations.

Repository infrastructure: We also intend to use this project to provide pathways and infrastructure to facilitate more collaborative work between librarians, faculty members, IT, and digital scholars at FIU and UF. Both institutions host digital collections on the SobekCM platform, an open-source repository software solution developed at UF. SobekCM offers strong functionality
for viewing and downloading newspapers and other materials at the item, issue or page level; however, it does not currently provide tools for users to download collection or title-level groups of files simultaneously or to easily access OCR text files. This project will help us document potential routes for making data discoverable through Sobek; in the near future, we will rely on FIU’s Research Data Portal, a local instance of Harvard’s Dataverse software, as a separate data repository space for this project. This will allow for branding, contextualization of the data sets, metadata, and mass downloads of the data. This makes for a more streamlined and easy user experience for researchers looking for this specific data, who may not want to sift through other datasets to find what they need.

**Expertise and training:** dLOC follows a model of shared governance, decentralized digitization and distributed collection development, thus giving Caribbean institutions, and those that know the collections most intimately, an important role in the decision making and production process. Recent professional development opportunities such as the 2019 NEH-funded institute, “Migration, Mobility, and Sustainability: Caribbean Studies and Digital Humanities,” have fostered potential collaborations and experimentation with data visualization and publication, particularly in pedagogical contexts. However, dLOC partners and affiliated researchers are often working in their own disciplinary or institutional silos as they seek ways to reuse collections, and few dLOC partners have sufficient capacity in data curation or analysis at scale. Project funding will allow not only the release of newspaper data, but more importantly focused collaboration with expert advisors and the launch of a community engagement effort that brings people together for events and trainings.

Project funding would also allow for the participation of Boyuan Guan, the Lead Developer for the FIU Libraries’ GIS Center and Digital Collections Center, who will collaborate on named entity extraction, geocoding, and data modeling as described further in the Draft Implementation Model component. He specializes in the development of web-based databases and GIS applications, programming, use of GIS software in civil engineering project management, digital repository systems, and metadata engines. This initiative will provide new opportunities for collaboration between the Digital Collections Center and GIS Center with dLOC. Additionally, Guan’s background in transportation engineering and computer science may lead to unanticipated insights. Guan’s expertise will ensure stronger technical documentation and provide team training to build capacity for text analysis and geocoding.

**Significance & Research Value**
Archival materials about hurricanes and tropical storms have become increasingly important for scholars of the Caribbean. People in the Caribbean have been coping with hurricanes and tropical storms for thousands of years. Hurricane data has been able to provide insight into cultural, economic, social, and environmental histories of the Caribbean. With limited resources, researchers have been able to provide documentation about struggles over disaster capitalism, labor, land, and climate. This data would contribute further evidence about the reported strength of storms and about
the resiliency of Caribbean people and society. Through this project, our scholars may investigate questions such as: How might we recover stories of individual people and places impacted by hurricanes across the region over time?; How has regional newspaper coverage of hurricanes over time compared with government reports?; and how might we track the locations of lesser known events where little data exists?

Methodologically, the project also responds to Northeastern University’s 2018 report on OCR research, which recommends that projects “create datasets with significant language variation and mixtures of languages and scripts.” Minimal geographic metadata and linguistic variation pose challenges as we seek to establish relationships between coverage of hurricanes, where they took place, and who was affected. One research question to be addressed with both technical and scholarly experts on the team will be the best way to identify and structure multilingual references to multiple entities associated with a single storm system, sometimes across a large geographic region. Documentation of this undertaking will be of interest to any researchers seeking to analyze disasters, social and cultural movements, or other events across national and linguistic boundaries.

(6) Project Plan

Ethical Considerations

The ethical development and reuse of digital collections through a shared governance model is central to dLOC’s institutional framework and day-to-day work. In this proposal, we are specifically addressing two areas where access to dLOC data would promote more equitable approaches to research and teaching:

1. There is a clear disparity regarding access to Caribbean newspapers; while digitized page images and in most cases searchable text are available, this project would begin to facilitate the kinds of research currently feasible with sources such as Chronicling America. This might include tracing social and reprinting networks, better understanding coverage of events with a regional impact, etc.

2. Our thematic focus on hurricanes and tropical cyclones will also address several ethical issues. While there is an ethical imperative to make more data about these disasters available to support research into Caribbean history, climate change, and other fields, we also acknowledge the ways in which reporting and research on disasters can omit individual and community identities or treat hurricane data as an extractable resource. Inspired by the Colored Conventions Project Principles, we will similarly seek to name specific people and places as a way to affirm their value and experience. Our collaboration with experts across disciplines will help ensure this data is contextualized thoughtfully.

The project will also focus on modeling better practices in fostering professional growth for all team members and acknowledging all contributions. This includes an emphasis on ensuring a positive graduate student internship experience based on UF’s long standing internship program, which
builds in supports for a living wage, resume development, and professional networking. At minimum, we will credit all who contribute over the course of the grant period and beyond on the project website, with a brief description of their role; we may also seek to attribute more granular contributions to project data or event organization as feasible.

**Draft Use Model**

**Leverage existing networks:** dLOC is a long-standing organization that reaches a range of distinct communities across disciplines and geographies. Members of the project team routinely attend conferences or engage with virtually connected networks of dLOC users, offering a sustainable and consistent set of opportunities to promote access to data over time and to continuously assess and iterate based on communities’ needs. Funding will allow for dedicated staff time and additional outreach opportunities during the grant period, but outreach will continue long-term.

**Promoting community engagement & interpretation:** Building sub-communities engaged with newspaper research and with hurricane or disaster studies is a crucial component in enhancing data and encouraging its use. During the latter half of the grant period, we will pilot a model to provide small stipends to institutions willing to organize events focused on experimenting with newspaper data. While we anticipate many of these will prioritize scholars and teachers in higher education as a primary audience, they may also include students, GLAM professionals, and participants in local history or civic engagement initiatives. Depending on local interests, these may feature a hands-on training session on using visualization software; an edit-a-thon to help document relationships between identifiable hurricanes, people, and places; or an opportunity to experiment with bulk data from our project alongside newspaper data from other regions. This series of events will seed efforts to demonstrate dLOC’s potential as a source for newspaper data and for better understanding the impact of and responses to hurricanes at local and regional levels.

To facilitate long-term sustainability and to strengthen components of this work, during and immediately following the grant period we will undertake assessment in the form of event evaluations and virtual town halls to get feedback on how community engagement efforts meet or do not meet specific needs. For instance, what additional documentation would make the data more useful to students or newcomers to the digital humanities? What barriers remain to conducting live events with regard to technical infrastructure or bandwidth? Ideally, we will be able to work with local event organizers to develop a “menu” of successful programs that can be replicated in the future with very little or no funding. Some of these may be dependent specifically on hurricane-focused data, while others may be adaptable to Caribbean newspaper data more broadly. Some may require external support such as software training from a dLOC community member, while others may be entirely self-guided. We are confident that we will be able to engage users in an ongoing conversation to determine what will be genuinely useful; for instance, recent virtual conversations with participants in dLOC’s year-long NEH institute have attracted consistent participation and concrete suggestions for building capacity in the field.
dLOC’s presence outside higher education, including professional development trainings at the Miami-Dade Public Schools and participation at the Miami Book Fair, provides additional long-term possibilities for broadening the audience and framing components of the project from a public humanities perspective. For instance, once we have a better understanding of the data we may be able to identify and produce simple data stories to augment existing online lesson plans available through the Florida & Puerto Rican Newspapers Project.

Encouraging adaptation of use model: Other users may include digital scholars, librarians, etc. interested in our approach to user community engagement and in developing thematic data toolkits in areas outside Caribbean Studies. As laid out in the Documentation Overview below, the project will develop training materials and workflows for others seeking to replicate this model, with release of user personas and any materials developed for or by local community events. We are particularly interested in reaching other organizations or communities that approach digital collecting building and analysis from a networked, shared governance perspective similar to dLOC’s. For instance, at the conclusion of the grant period we plan to organize a virtual outreach event specifically targeted toward state or regional aggregations (including but not limited to DPLA hubs) and other community-driven collections (e.g. South Asian American Digital Archive, Advanced Research Consortium nodes).

Other model projects: We will produce a summary of other projects with outreach strategies that prioritize strong documentation and community engagement, particularly in the context of newspaper data and named entity extraction. In particular, this includes the Linked Jazz project and related Semantic Lab at the Pratt Institute, which seeks out community collaboration in enhancing relationship data; and the Viral Networks project based at Virginia Tech, which has built upon an earlier newspaper analysis effort focused on epidemiology and sustained engagement for nearly a decade through symposia, hands-on workshops, and a recent publication.

Positions and duties: The project co-leads and advisory committee will play a crucial role in seeking feedback and in facilitating professional development and outreach to current dLOC constituencies and to other potential user communities (e.g. digital humanists with an interest in newspaper research; public scholars developing narratives around hurricane and disaster research). Technical experts in metadata, text analysis, and data visualization will also support use by making sure datasets are well-described, discoverable, and citable, and by creating training materials directed at novice users. The project co-leads will also focus on ensuring that the use model is adaptable by other institutions by documenting and disseminating lessons learned--in international collaboration, project marketing, and data documentation.

Positions & Duties Summary

<table>
<thead>
<tr>
<th>Use-Focused Responsibility</th>
<th>Primary Individual/Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online community (web, social media,</td>
<td>Project leads with support from all</td>
</tr>
<tr>
<td>Google group)</td>
<td>team members</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Conference outreach</td>
<td>Project leads</td>
</tr>
<tr>
<td>Data documentation/ thematic finding aid</td>
<td>Project leads; FIU technical team and advisory committee</td>
</tr>
<tr>
<td>Develop or adapt training guides for analyzing newspaper data and for adapting new thematic toolkits</td>
<td>Project leads; FIU technical team</td>
</tr>
<tr>
<td>Local community training events and edit-a-thons</td>
<td>Project leads; advisory committee; community partners</td>
</tr>
</tbody>
</table>

**Sustaining the use model:** While routine outreach at conferences should help ensure continued awareness and use of plug-and-play datasets and some experimentation with newspaper data, it will be more challenging to launch and sustain development of future thematic data toolkits and other interpretive research projects. To alleviate this concern, one goal of the project will be to collaborate with the advisory committee and to document a network of scholarly and technical experts—in hurricane and disaster studies, in textual and image data analysis, and in humanities data curation—who future users may be able to call upon. The dLOC network also provides members with support for grant development, a service that could help foster future investment in projects leveraging newspaper data.

**Draft Implementation Model**

**Phase 1: Newspaper data processing & dissemination**

The first step will focus on finalizing a list of newspaper titles for inclusion in the project. We will aim to make available approximately 200,000 page images and OCR text for bulk download based on the following criteria:

- Contribution to breadth across geographic, linguistic, and temporal coverage.
- Availability of acceptable OCR text. This may privilege newspapers that have been digitized more recently or where OCR processing has recently been redone (e.g. for many of dLOC’s Cuban newspaper titles). Born-digital titles may also be considered.
- For titles in need of OCR reprocessing, we will seek out publications that contain a small number of issues. This will enable UF’s digital collections team to complete OCR for all issues within the publication. We will reprocess OCR only where the results are most likely to be acceptable (e.g. high contrast, consistent layout, minimal deterioration).
- Titles must not already be available in Chronicling America, Europeana Newspapers, or other sources where bulk download or computational analysis is currently feasible.
- Permission must have been previously granted by the copyright holder for noncommercial use, or titles must be in the public domain.
To make the data available for download, we will complete the following steps:

1. Where necessary, undertake OCR reprocessing and quality assurance and ingest updated text files in dLOC. Work with advisory committee to spot check records across text in multiple languages. (UF technical team; advisory committee)

2. To facilitate UI download access to batch data: Copy available metadata and file directories to public FIU data repository. Directories are currently structured in pairtree format, also used by resources such as HathiTrust. Data packages will contain page images in JPEG 2000 and PDF formats; uncorrected OCR text with word bounding boxes; and structural metadata. TIFF files will not be disseminated as part of data packages. (UF & FIU technical teams)

3. To facilitate access to bulk text-only downloads: For each publication, extract and package OCR text files to provide bulk download access at title level outside nested directory structure. Each download will include a file manifest with page identifiers and corresponding issue dates. These files will be made available both through the dLOC and FIU data repository interfaces. Each data set will be assigned its own DOI, which allows us to link bi-directionally from the newspaper collection repository to the associated data. (UF & FIU technical teams)

4. In addition to the metadata which will accompany each of the publication data packages, the Dataverse repository will also contain information about the corpus of materials included in this project as well as pertinent documentation about the preparation and structure of the materials for bulk download.

**Phase 2: Derive thematic datasets**

To highlight the value of a collections as data approach and to open up dLOC newspaper data to a larger audience, we will develop a pilot thematic toolkit that provides structured data focused on newspaper coverage of hurricanes and tropical cyclones. As much as possible, we will undertake this phase with a focus on developing well-documented, replicable workflows for application with other themes (see Documentation Overview below). This will require the following steps:

1. Collaboratively develop controlled vocabulary comprised of a target keyword list and tagging, which will include all references to “hurricane” and related terms across English, Spanish, Dutch, and French, as well as other languages where feasible. Also include proper names assigned to particular storms (e.g. Andrew; San Lorenzo). (Scholarly lead; advisory committee).

2. Create corpus made up of OCR text from Phase 1 as well as Chronicling America data for at least one Puerto Rican newspaper, both to ensure geographic coverage and to test interoperability between data sources as a major goal of this proposal. (FIU technical team)

3. Create a simple topic classifier model to automatically generate new tags within the dataset in order to expand the list originally defined by the scholarly lead and advisory committee. Define a named entities list. Develop a text extraction model to automatically generate additional keywords based on this list. Compile all references to targeted keywords and export each section of text containing a relevant term to tab delimited files and to CSV with
identifier and corresponding page URI. (FIU technical team; multi-lingual advisory committee members)

4. Experiment with additional named entity extraction tools (e.g. Stanford NLTK) as well as manual methods to geolocate hurricane references and to identify specific individuals or organizations co-located with references to hurricanes, referring to existing name authorities wherever possible. Develop data model for storing and disseminating entity and relationship information in appropriate formats (CSV, JSON-LD, etc.). (Project leads; FIU technical team; multi-lingual advisory committee members)

5. Archive the derived thematic datasets as a subsection of the data repository containing the content for bulk download. These thematic datasets will include detailed documentation of the processes for identifying target keywords, results and recommendations for further work to improve OCR quality and data interoperability, and process for identifying entity and relationship information.

This phase is crucial to making the data actionable for our user community. It will present intellectual and technical challenges in name disambiguation, translation, and data modeling; however, it will also present opportunities to better articulate relevant research questions and the ways in which dLOC’s technical partners can support that research. Our primary goal is to collaborate to build capacity across the dLOC network for developing documentation, templates, and training opportunities during and after the grant period.

Phase 3: Develop thematic toolkit and documentation
As a final step in making data accessible, we will prepare a thematic data toolkit that will include a snapshot of the hurricane dataset(s); a descriptive document or finding aid with information about toolkit data assets; and an inventory of other relevant digital objects in dLOC (e.g. photographs, government reports) and/or external datasets (e.g. NOAA hurricane data).

Additionally, the toolkit will include 4-5 short tutorials demonstrating specific ways to analyze or visualize the data and appropriate tools to explore, based on documented processes and results from the project team’s initial data exploration. These will be directed toward novice or intermediate users and will include step-by-step walkthroughs for tasks such as associating entities across two or more datasets; preparing and interpreting data when using tools such as Palladio; or even testing our target keyword list with software such as AntConc as one step toward understanding how the thematic data was created. Additionally, drawing on projects like DataBasic, we will ensure that our tutorials help users better understand how to formulate questions out of the data, and how to identify the stories that the data can tell.

These guides will be crucial to contextualizing the data for students in the digital humanities or other fields as they begin to understand both the potential and the limitations of structured data. Wherever possible, we will adapt guides from other sources such as Programming Historian and will make these available in English, Spanish, and French. Depending on timing, these tutorials may
offer a starting point for local user community events, but they should also be informed by those events as specific needs and suggestions emerge.

**Encouraging replication of implementation model:** Phase 1 of the project will be most adaptable to other institutions and collections, as it relies on expertise and technical infrastructure common to a range of cultural institutions. For instance, UF is already planning to incorporate workflows refined through this project into future digitization initiatives across collections. Our focus on newspapers is likely to be of wide interest, especially to those institutions that for a variety of reasons are not included in Chronicling America or that are grappling with legacy OCR. In addition, the use of an established open access data repository software with well established standards will provide a low barrier for others to replicate and implement a similar model of access, sharing, discoverability, and storage. While our pilot thematic toolkit and derived datasets are more specialized and may not be fully scalable, we plan to create a guide that documents our general methods for creating such a toolkit and recommendations to others regardless of subject matter. This guide will outline processes such as (1) identifying potential source data; (2) reviewing and remediating source data; (3) data storage/archiving; (4) identifying potential research questions/application of data; (5) creating derivative data sets; (6) methodologies and tools for analysis.

**Other model projects:** For data preparation and dissemination, we will look most closely to the approach Chronicling America has taken in making both textual OCR and page data available, and we will include information in our documentation for those who wish to develop corpora containing papers from both data sources. This will include consulting internal documentation models for the Florida and Puerto Rican Newspaper Project. Europeana Newspapers NER provides a rich example as we consider feasible methods for identifying events, people, and locations within our own corpus. In curating derived datasets and providing interpretive context, we will seek to emulate approaches such as that described in Katie Rawson and Trevor Muñoz’s 2016 article, “Against Cleaning,” which offers a general framework for making data curation decisions explicit and balancing data normalization with preservation of data diversity and even inconsistency. This approach is crucial as we seek to provide datasets that are meaningful and actionable without making opaque the inherent complexity of newspaper coverage over time, language, and space.

**Positions and duties:**
Implementation will rely heavily on team members with technical expertise in digital production, OCR workflows, repository infrastructure, metadata, and text analysis. While our project brings together this expertise from across institutions, the staffing model is readily adaptable for single institutions with sufficient expertise in across these areas. Training on both a local and broader community level will be key to ensuring the sustainability of our approach. Depending on interest and expertise, the graduate intern will have an opportunity to contribute throughout the data curation
### Positions & Duties Summary

<table>
<thead>
<tr>
<th>Implementation-Focused Responsibility</th>
<th>Primary Individual/Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project management and coordination</td>
<td>Collins (project-wide and UF); Rogers (FIU)</td>
</tr>
<tr>
<td>Identifying and preparing data for dissemination, including OCR quality control</td>
<td>UF/FIU digital collections and metadata experts, with support from multi-lingual members of advisory committee</td>
</tr>
<tr>
<td>Packaging, formatting, and describing data</td>
<td>UF/FIU digital collections, repository, and metadata experts</td>
</tr>
<tr>
<td>Developing corpus of Caribbean newspaper hurricane coverage</td>
<td>Castro; Guan; project leads; graduate intern; support from multi-lingual members of advisory committee</td>
</tr>
<tr>
<td>Experimentation with semi-automated and manual methods for extracting named entities associated with hurricane coverage, including data modeling</td>
<td>Guan; Castro; Bakker; project leads; graduate intern; support from multi-lingual members of advisory committee</td>
</tr>
<tr>
<td>Create thematic toolkit, including derived datasets, documentation, training modules, and references to other relevant data</td>
<td>Scholarly lead; project lead; project intern</td>
</tr>
<tr>
<td>Provide “train the trainers” workshop and documentation to build local capacity for text analysis and data modeling</td>
<td>Guan; Castro; Bakker</td>
</tr>
<tr>
<td>Create training materials for those seeking to adapt implementation model for other collections</td>
<td>Project leads; technical partners</td>
</tr>
</tbody>
</table>

### Data Overview

To summarize, data to be disseminated will include the following:

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Format(s)</th>
<th>Source(s)</th>
<th>Access Point(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caribbean Newspapers page images</td>
<td>JPEG2000/PDF</td>
<td>dLOC/UF</td>
<td>FIU Dataverse; page-level access in dLOC</td>
</tr>
<tr>
<td>Caribbean Newspapers OCR</td>
<td>ALTO/XML; TXT</td>
<td>dLOC/UF</td>
<td>FIU Dataverse; potential for</td>
</tr>
<tr>
<td>text</td>
<td>Structural and descriptive metadata</td>
<td>METS XML</td>
<td>dLOC/UF</td>
</tr>
<tr>
<td>------</td>
<td>------------------------------------</td>
<td>----------</td>
<td>---------</td>
</tr>
<tr>
<td></td>
<td>Dataset including references to hurricanes and related personal or geographic entities</td>
<td>CSV; JSON-LD</td>
<td>Derived from Caribbean Newspaper data and other relevant titles currently in Chronicling America</td>
</tr>
<tr>
<td></td>
<td>Scripts</td>
<td>Python</td>
<td>FIU Technical Team</td>
</tr>
</tbody>
</table>

**Documentation Overview:**
The following summarizes documentation to be generated over the course of the grant period, as described above. The Project Lead (Collins) and Project Intern will be responsible for assigning documentation tasks and collecting and depositing outputs. Materials will be made available through dLOC, GitHub, Dataverse, and the project website as appropriate. Materials will be shared under a Creative Commons Attribution-Non Commercial 4.0 International License except in cases where local community event organizers choose a different license for their outputs.

<table>
<thead>
<tr>
<th>Item(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project website with overview, monthly updates, and resources</td>
</tr>
<tr>
<td>Statement of ethical principles for collaboration and data reuse</td>
</tr>
<tr>
<td>Environmental scan to contextualize use and implementation models</td>
</tr>
<tr>
<td>Advisory board meeting minutes</td>
</tr>
<tr>
<td>Evaluation of OCR quality, including results and template for review at distinct points (initial title selection, recognition of target keywords, named entity recognition)</td>
</tr>
<tr>
<td>Workflows and scripts for migrating data from UF to FIU repository and for extracting plaintext files for deposit in dLOC</td>
</tr>
<tr>
<td>Data citation metadata, discipline-specific metadata and file level documentation</td>
</tr>
<tr>
<td>Workflows for creating dataset from Caribbean Newspapers and Chronicling America (Puerto Rico) titles and any interoperability obstacles</td>
</tr>
<tr>
<td>Derived data documentation (including target keyword list and methodology; workflows and scripts for extracting and structuring hurricane-related data; key obstacles and</td>
</tr>
<tr>
<td>Troubleshooting; data model</td>
</tr>
<tr>
<td>-------------------------------</td>
</tr>
<tr>
<td>Project team training materials focused on text analysis and geocoding</td>
</tr>
<tr>
<td>Community event materials (including call for participation; marketing resources; slides/recordings; and assessment instruments)</td>
</tr>
<tr>
<td>Conference presentation slides/recordings</td>
</tr>
<tr>
<td>Thematic data documentation (including target keyword list; plain language summary of methodology; descriptive “finding aid” to facilitate discovery of underlying newspaper files and other relevant datasets; basic tutorials)</td>
</tr>
<tr>
<td>Virtual outreach/webinars to promote use and implementation models</td>
</tr>
<tr>
<td>Project report outlining summary of methods, accomplishments, and outputs (including use/implementation models), as well as lessons learned and future goals. Areas of likely broad interest will include OCR evaluation; ethics of collections as data in a shared governance environment; and our hurricane-focused thematic approach.</td>
</tr>
</tbody>
</table>

**Stewardship & Sustainability:**

After refining workflows and points of data discovery as described in Phase 1, UF and FIU are committed to adopting these processes long-term for future digitization initiatives as well as existing collections. UF has already begun reprocessing legacy OCR, and this project will help prioritize future titles slated for reprocessing through collaboration with FIU’s digital collections team and by testing OCR quality in research contexts beyond basic text search.

For Phases 2 and 3, we will consider sustainability from two perspectives, for the hurricane-focused thematic toolkit and for the broader data analysis and toolkit models:

- For the former, project team members will commit at minimum to ensuring toolkit assets—including derived datasets and documentation—are available through dLOC and the FIU data repository for long-term preservation. Data will be stored and served in FIU’s centralized computing framework, a cloud computing infrastructure with 22 servers, over 220 TB storage space, and sufficient redundancy. Files will be routinely backed up on a weekly schedule, with versioning and a disaster recovery (DR) setup located in Tallahassee, Florida. As resources allow and as more dLOC newspaper data is made publicly available, we will continue to grow these datasets to include additional references to hurricanes and impacted people and locations.

- To enable the project team and other researchers to replicate the thematic toolkit model, we will also maintain workflow documentation, data modeling guidance, relevant training materials, and project code for long-term access and preservation in dLOC and GitHub.

Finally, we will write and disseminate a project white paper and other publications as appropriate to document project goals, lessons learned, and use cases.
### Timeline of completion

<table>
<thead>
<tr>
<th>Quarter 1 (Jan.-Mar. 2020)</th>
<th>Activities &amp; Responsible Team Member(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>● Review and finalize titles for inclusion (Project leads, partners, advisory committee)</td>
</tr>
<tr>
<td></td>
<td>● Undertake OCR as necessary and implement quality assurance (UF technical partners; multi-lingual members of advisory committee)</td>
</tr>
<tr>
<td></td>
<td>● Hire graduate student intern (Project lead)</td>
</tr>
<tr>
<td></td>
<td>● Advisory committee virtual meeting</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quarter 2 (April-June 2020)</th>
<th>Activities &amp; Responsible Team Member(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>● Prepare bulk download packages and disseminate via FIU Dataverse and dLOC (UF and FIU technical partners)</td>
</tr>
<tr>
<td></td>
<td>● Launch project website with documentation for data access (Project leads)</td>
</tr>
<tr>
<td></td>
<td>● Develop target keyword list for hurricane thematic text analysis (Scholarly lead; advisory committee)</td>
</tr>
<tr>
<td></td>
<td>● Advisory committee virtual meeting</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quarter 3 (July-Sept. 2020)</th>
<th>Activities &amp; Responsible Team Member(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>● Refine FIU Dataverse access and determine appropriate points of discovery via other platforms (FIU technical partners)</td>
</tr>
<tr>
<td></td>
<td>● Keyword analysis and data extraction (FIU technical team; multi-lingual members of advisory committee)</td>
</tr>
<tr>
<td></td>
<td>● Collaborate to implement text analysis and to develop thematic data model (Project leads, technical partners, graduate intern)</td>
</tr>
<tr>
<td></td>
<td>● Use topic classifier model to extract geographic or personal named entities where feasible and associate with existing authorities (VIAF, OpenStreetMaps) (Project leads, technical partners, graduate intern)</td>
</tr>
<tr>
<td></td>
<td>● Release call for local community events (Project lead, graduate intern)</td>
</tr>
<tr>
<td></td>
<td>● Disseminate information about project and initial findings on project website and conference presentations (Project lead, scholarly lead, graduate intern)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quarter 4 (Oct.-Dec. 2020)</th>
<th>Activities &amp; Responsible Team Member(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>● Continue data processing and correction of major errors (Project lead, FIU technical team, graduate intern)</td>
</tr>
<tr>
<td></td>
<td>● Finalize thematic data toolkit with derived datasets, documentation, and use case examples (Project lead, scholarly lead, graduate intern)</td>
</tr>
<tr>
<td></td>
<td>● Announce local community events and work with</td>
</tr>
<tr>
<td>Task</td>
<td>Responsible Party</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Organizers to identify specific data or training needs</td>
<td>Project lead, Scholarly lead, graduate intern</td>
</tr>
<tr>
<td>• Conference outreach</td>
<td>Project leads, graduate intern</td>
</tr>
<tr>
<td>• Complete and release documentation for technical workflows and</td>
<td>Project lead, FIU technical team</td>
</tr>
<tr>
<td>associated code</td>
<td></td>
</tr>
<tr>
<td>• Advisory Committee virtual meeting</td>
<td></td>
</tr>
<tr>
<td>Quarter 5 (Jan.-Mar. 2021)</td>
<td></td>
</tr>
<tr>
<td>• Provide facilitation and follow-up support to local community</td>
<td>Project leads, FIU technical team</td>
</tr>
<tr>
<td>event organizers</td>
<td></td>
</tr>
<tr>
<td>• Virtual training opportunities</td>
<td>Project leads; advisory committee</td>
</tr>
</tbody>
</table>
Appendix A: List of Newspaper Titles

The titles below are candidates for inclusion in bulk data downloads made available. This list excludes titles from Puerto Rico and the Virgin Islands that are already accessible or will soon be accessible through Chronicling America.

Abaconian (Bahamas, 1993-present) https://dloc.com/UF00093713/00001/allvolumes

Carteles (Cuba, 1919 - 1927) - https://dloc.com/AA00065193/00001

Le Civilisateur (Haiti, 1870-1873) - https://dloc.com/AA00062914/00001/allvolumes

Haïti Illustrée (Haiti, 1890-1892) - https://dloc.com/AA00062728/00001/allvolumes

Listin Diario (Dominican Republic), 1909-1930 - https://www.dloc.com/AA00021654/00006/allvolumes

Outlook (Belize, 1945-1946) - https://dloc.com/AA00064484/00001/allvolumes


Appendix B: Caribbean Data Curation Graduate Intern Position Description

Term(s): Summer/Fall 2020

Compensation: $15/hr for up to 480 hrs

Position overview: Reporting to the Project Lead/Scholarly Communications Librarian at the University of Florida, the Caribbean Data Curation Graduate Intern will play a meaningful role in developing the grant-funded initiative dLOC as Data: A Thematic Approach to Caribbean Newspapers. In partnership with the Digital Library of the Caribbean (dLOC), this project seeks to enhance access to digital newspapers from the Caribbean by making newspaper “data”-- including text and page images--available for new kinds of scholarly and educational exploration. Additionally, the project will result in a toolkit focusing on ways newspaper data can offer insights into the history and impact of hurricanes and tropical storms in the Caribbean. Based on interest and experience, the intern will collaborate with team members at UF and Florida International University to help prepare and disseminate data, to engage with dLOC community partners, and to create training and demonstration materials. The project team is committed to supporting the intern’s professional goals and to acknowledging all contributions.

Summary of duties: The intern will work alongside team members to create a thematic toolkit focused on newspaper coverage of hurricanes in the Caribbean. This may include participation in gathering and enhancing data; providing documentation to help others understand how to use the data; and identifying other datasets or primary sources relevant to this topic. The intern will also play a key role in outreach, including creating blog and social media content and responding to inquiries from community partners. Duties will be finalized at the beginning of the internship to align with the intern’s professional goals and desired areas of experience. All Smathers Libraries interns are required to participate in a CV writing workshop and to give a public presentation on their work. The intern will also be invited to participate in a 1-day project meeting and training session to be hosted at FIU in Miami.

Required qualifications:
1. Enrollment in a relevant advanced degree program at the University of Florida. Many fields may be considered relevant; candidates should describe why their academic background supports the position duties in the letter of interest.
2. Interest in the digital humanities and a willingness to experiment with new technologies.
3. Experience giving presentations and teaching others in formal or informal settings.
4. Strong written communication skills and experience writing for public audiences.
5. Enthusiasm for collaborating with international partners.
6. Experience editing basic websites and familiarity with Microsoft Excel or other spreadsheet applications.
Preferred qualifications:

Note that the following are preferred, not required, and strong candidates need not meet every qualification.

1. Some knowledge of text analysis and data visualization concepts and software and experience preparing data for analysis.
2. Reading knowledge of French and/or Spanish.
3. Experience giving presentations or trainings in an online setting.
### Collections as Data Budget: Florida International University

<table>
<thead>
<tr>
<th>Details/Notes</th>
<th>Yr. 1 (1/1/20-12/31/20)</th>
<th>Yr. 2 (1/1/21 - 3/31/21)</th>
<th>Grand Total (1/1/20-3/31/21)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Salaries &amp; Wages</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scholarly Lead (St. Hubert)</td>
<td>%5 $65,000</td>
<td>$3,250.00</td>
<td>$4,062.50</td>
</tr>
<tr>
<td>Administrative Lead (Asencio)</td>
<td>%2 $57,424</td>
<td>$861.36</td>
<td>$1,076.69</td>
</tr>
<tr>
<td>Digital Collections (Rogers)</td>
<td>%5 $67,367</td>
<td>$3,368.35</td>
<td>$4,210.40</td>
</tr>
<tr>
<td>GIS (Guan)</td>
<td>%4 $80,566 (12 months total)</td>
<td>$2,416.98</td>
<td>$3,222.64</td>
</tr>
<tr>
<td>Digital Collections (Bakker)</td>
<td>%5 $46,500</td>
<td>$2,325.00</td>
<td>$2,906.25</td>
</tr>
<tr>
<td>Digital Humanities (Castro)</td>
<td>%4 $45,000</td>
<td>$1,890.00</td>
<td>$2,396.25</td>
</tr>
<tr>
<td>Repository Coordinator (Krefft)</td>
<td>%3 $54,025</td>
<td>$1,620.75</td>
<td>$2,025.93</td>
</tr>
<tr>
<td>Advisor (Capo)</td>
<td>%1 $85,000</td>
<td>$500.00</td>
<td>$500.00</td>
</tr>
<tr>
<td><strong>2. Fringe Benefits</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scholarly Lead (St. Hubert)</td>
<td>34.01%</td>
<td>$1,547.45</td>
<td>$1,823.78</td>
</tr>
<tr>
<td>Administrative Lead (Asencio)</td>
<td>34.01%</td>
<td>$292.94</td>
<td>$366.17</td>
</tr>
<tr>
<td>Digital Collections (Rogers)</td>
<td>34.01%</td>
<td>$1,145.57</td>
<td>$1,431.95</td>
</tr>
<tr>
<td>GIS (Guan)</td>
<td>34.01%</td>
<td>$820.86</td>
<td>$1,094.86</td>
</tr>
<tr>
<td>Digital Collections (Bakker)</td>
<td>34.01%</td>
<td>$790.73</td>
<td>$988.41</td>
</tr>
<tr>
<td>Digital Humanities (Castro)</td>
<td>34.01%</td>
<td>$642.78</td>
<td>$814.95</td>
</tr>
<tr>
<td>Repository Coordinator (Krefft)</td>
<td>34.01%</td>
<td>$551.21</td>
<td>$889.01</td>
</tr>
<tr>
<td>Advisor (Capo)</td>
<td>34.01%</td>
<td>$170.04</td>
<td>$170.04</td>
</tr>
<tr>
<td><strong>3. Consultant/collaborator Fees</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advisor honorarium (Gil)</td>
<td></td>
<td>$500.00</td>
<td>$500.00</td>
</tr>
<tr>
<td>Advisor honorarium (Levi)</td>
<td></td>
<td>$500.00</td>
<td>$500.00</td>
</tr>
<tr>
<td>Advisor honorarium (Mohan)</td>
<td></td>
<td>$500.00</td>
<td>$500.00</td>
</tr>
<tr>
<td><strong>4. Travel</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conference 1 - Digital Library Federation (DLF)</td>
<td>From Grant - $2,011.17. FIU Cost Share Total: $1,092.83.</td>
<td>$3,104.00</td>
<td>$3,104.00</td>
</tr>
<tr>
<td>Conference 2 - Caribbean Digital</td>
<td></td>
<td>$1,500.00</td>
<td>$1,500.00</td>
</tr>
<tr>
<td><strong>5. Supplies &amp; Materials</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Community event stipends</td>
<td>$300 available to up to 10 institutions interested in hosting local edit-a-thons/hackathons</td>
<td>$3,000.00</td>
<td>$3,000.00</td>
</tr>
<tr>
<td>Data Storage Cost</td>
<td>FIU Cost share - for the JPEG200 and PDF files to link text data back to the dLOC repository</td>
<td>$2,200</td>
<td>$2,200</td>
</tr>
<tr>
<td>6. Subawards</td>
<td>See below</td>
<td></td>
<td>$14,209.00</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
<td>---</td>
<td>----------</td>
</tr>
<tr>
<td>a. Requested funding</td>
<td></td>
<td></td>
<td>50,000</td>
</tr>
<tr>
<td>b. Cost sharing</td>
<td></td>
<td></td>
<td>$9,905</td>
</tr>
<tr>
<td>Total project funding</td>
<td></td>
<td></td>
<td>59,905</td>
</tr>
</tbody>
</table>
### 1. Salaries & Wages

<table>
<thead>
<tr>
<th>Details/Notes</th>
<th>Year 1 (1/1/20-12/31/20)</th>
<th>Year 2 (1/1/21-3/31/2021)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project lead (Collins)</td>
<td>5%</td>
<td>$3,666.00</td>
<td>$923.00</td>
</tr>
<tr>
<td>Caribbean Data Curation Graduate Intern</td>
<td>$15x480 hrs</td>
<td>$7,200.00</td>
<td>$7,200.00</td>
</tr>
<tr>
<td>Digital Support Services (Dinsmore)</td>
<td>2% (cost share)</td>
<td>$1,854.00</td>
<td>$1,854.00</td>
</tr>
<tr>
<td>Digital Production (Perry)</td>
<td>2% (cost share)</td>
<td>$1,139.00</td>
<td>$1,139.00</td>
</tr>
<tr>
<td>Digital Partnerships (Taylor)</td>
<td>2% (cost share)</td>
<td>$2,140.00</td>
<td>$2,140.00</td>
</tr>
</tbody>
</table>

### 2. Fringe Benefits

<table>
<thead>
<tr>
<th>Details/Notes</th>
<th>Year 1 (1/1/20-12/31/20)</th>
<th>Year 2 (1/1/21-3/31/2021)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project lead (Collins)</td>
<td>26.80%</td>
<td>$982.00</td>
<td>$247.00</td>
</tr>
<tr>
<td>Caribbean Data Curation Graduate Intern</td>
<td>5.70%</td>
<td>$410.00</td>
<td>$410.00</td>
</tr>
<tr>
<td>Digital Support Services (Dinsmore)</td>
<td>26.8% (cost share)</td>
<td>$497.00</td>
<td>$497.00</td>
</tr>
<tr>
<td>Digital Production (Perry)</td>
<td>35.7% (cost share)</td>
<td>$407.00</td>
<td>$407.00</td>
</tr>
<tr>
<td>Digital Partnerships (Taylor)</td>
<td>26.8% (cost share)</td>
<td>$575.00</td>
<td>$575.00</td>
</tr>
</tbody>
</table>

### 3. Consultant Fees

### 4. Travel

2-day travel (accommodations, per diem) for project lead and graduate intern

<table>
<thead>
<tr>
<th>Details/Notes</th>
<th>Year 1 (1/1/20-12/31/20)</th>
<th>Year 2 (1/1/21-3/31/2021)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Travel to workshop (FIU/Miami)</td>
<td></td>
<td>$781.00</td>
<td>$781.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Details/Notes</th>
<th>Year 1 (1/1/20-12/31/20)</th>
<th>Year 2 (1/1/21-3/31/2021)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Requested funding</td>
<td>$14,209.00</td>
<td></td>
<td>$14,209.00</td>
</tr>
<tr>
<td>b. Cost sharing</td>
<td>$6,612.00</td>
<td></td>
<td>$6,612.00</td>
</tr>
</tbody>
</table>

Total subaward funding $20,821.00
October 28, 2019

Collections as Data: Part to Whole
Cohort 2 Application

Dear Collections as Data: Part to Whole Grant Committee,

On behalf of the Digital Library of the Caribbean (dLOC) and Florida International University Libraries, I want to express support for the “dLOC as Data: A Thematic Approach to Caribbean Newspapers” grant application. Please acknowledge my commitment as Executive Director of the Digital Library of the Caribbean (dLOC) to collaborate with Senior Administrative Lead Jamie Rogers, and Project Leads Perry Collins and Hadassah St. Hubert. The project leads will have access to any resources they need during the project. They will be able to participate in our leadership meetings to update our members about the impact of their work. Dr. St. Hubert, as ex-Officio for the Scholarly Advisory Board, will have our support to disseminate findings to our various audiences. In addition, our office will support preparations for all virtual and in-person meetings as well as community engagement events.

The “dLOC as Data: A Thematic Approach to Caribbean Newspapers” project is important to the work we do because it helps us assess and begin to address some of the biggest challenges we face in providing accessible data to scholars and researchers. This project ultimately increases the accessibility of primary and secondary open-access research and education materials. The importance of this project to our project leads is further amplified by the responsibilities and roles they play on the frontlines of scholarly research and education endeavors.

dLOC is hosted through the SobekCM platform, an open-source repository software solution developed at the University of Florida and implemented by both UF and FIU for most of our digital collections. Providing increased functionality to the platform has been part of our on-going goals. Having access to OCR text and evaluating multilingual OCR quality, greatly increase knowledge about and from the Caribbean. Hurricanes and Tropical Storms have shaped our historical context in South Florida and the Caribbean. As the largest and busiest online Caribbean content library, we will make sure to leverage our scholarly board, our collaborative work within and across institutions, as well as dLOC site’s reach (www.dloc.com) to engage many about the availability of these data-sets and share the lessons we learn from implementing multilingual OCR.

To reiterate my commitment to Caribbean Studies, Open Access education, and research materials, I fully support and endorse the “dLOC as Data: A Thematic Approach to Caribbean Newspapers” project. Please feel free to contact me if you have questions or require additional information.

Sincerely,

Miguel Asencio
Director
October 18, 2019

To: Collections as Data: Part to Whole  
From: Jamie Rogers, Florida International University  
RE: dLOC as Data: A Thematic Approach to Caribbean Newspapers

As Assistant Director of Florida International University’s Digital Collections Center (DCC), I am pleased to submit this letter of support for the initiative dLOC as Data: A Thematic Approach to Caribbean Newspapers. This proposed project aims to provide ready access to a large corpus of Caribbean newspaper textual data as well as a prototype thematic toolkit focused on the impact and response to hurricanes and tropical cyclones across the Caribbean. The outcomes of this project have the potential to be far reaching, not only serving the community of students and scholars who study the history of the Caribbean and the impacts of natural disasters, they may also provide insights into future humanitarian efforts and disaster recovery.

Serving as co-administrative lead for the project, I will direct the efforts of FIU’s technical team in their execution of quality control of OCR outputs for the project’s newspaper data as well as archiving and preparation of the data for bulk download utilizing FIU’s Dataverse. The FIU technical team will also perform textual analysis, data extraction, and the development of curated thematic derivative data sets to be included in the toolkit, which will be archived, preserved, and made available to students and scholars.

I enthusiastically support this initiative as it uniquely addresses pressing concerns as climate change increasingly impacts our universities, local communities, and Caribbean neighbors. It is also a tremendous opportunity for professional development with the FIU Digital Collections Center, to expand the skill sets of our library team, who in turn support our students and faculty with expanded research capacities. This project will also serve as a pilot for future collections as data endeavors across both institutions and within dLOC. If our proposal is accepted, we will ensure our team has the necessary support and time allotted to accomplish the goals of this initiative.

Sincerely,

Jamie Rogers  
Assistant Director  
Digital Collections Center  
Florida International University
October 20, 2019

Dear Review Committee:

I am writing to express my commitment to participate as project lead for dLOC as Data: A Thematic Approach to Caribbean Newspapers, and to further describe my role in managing this important effort.

As a librarian within the UF Libraries’ Digital Partnerships & Strategies unit, a key part of my position focuses on developing infrastructure for emerging forms of publication and digital scholarship and pedagogy. One of my department’s major goals over the next two years will be to host and promote awareness of foundational platforms, data, and training opportunities for those who are eager to engage with collections held by the Libraries and our partners in new ways. This project will complement our broader digital scholarship program and demonstrate the potential for initiatives that move beyond digitization.

My grant-supported role as co-lead will allow me dedicated time to focus on project management over the course of the grant period. As outlined in the proposal implementation model and timeline to completion, I will participate in most steps of the project, with a more active role in liaising with the digital production team at UF and in supervising the graduate intern. All team members will document contributions via a project management platform such as Trello, and I will act as a primary source of communication to the advisory committee to ensure they have opportunities to participate in meaningful, well-defined ways. As project lead, I will oversee adherence to ethical practices as described in the narrative, ensuring all contributions are acknowledged.

I will also play a leadership role, along with the scholarly lead and graduate intern, as a community liaison in disseminating information about the project to dLOC and other stakeholders and in seeking input from new and current partners. This will include maintenance of a public website and discussion group and coordination of local events to experiment with newspaper data. Because of my broad knowledge of the field as a former program officer in the NEH Office of Digital Humanities, I am also well-positioned to reach out to networks beyond Caribbean studies to encourage adaptation of the project’s use and implementation models by other organizations—and to seek feedback on these models from other experts in the field.
Beyond the grant period, UF’s digital production and library technology colleagues have already expressed a willingness to refine current workflows and potentially make changes to our UF-built, open-source repository platform in order to enable a collections as data approach. This would help meet demand for access to dLOC textual and image data as well as data from collections such as the Baldwin Library of Children’s Literature, which is currently completing a planning grant to enable computational analysis. This potential for long-term sustainability and development will have a crucial and immediate impact on our partners and colleagues, especially those who rely on UF’s technical infrastructure in order to provide access to their own collections.

Sincerely,

Perry Collins
Scholarly Communications Librarian
University of Florida
352-273-2710
perrycollins@ufl.edu
October 28, 2019

Collections as Data: Part to Whole
From: Hadassah St. Hubert, Ph.D.
Cohort 2 Application: dLOC as Data: A Thematic Approach to Caribbean Newspapers

Dear Collections as Data: Part to Whole Grant Committee,

As the CLIR Postdoctoral Fellow in Data Curation for Latin American and Caribbean Studies at Digital Library of the Caribbean (dLOC), it is with great pleasure that I submit this letter to support the “dLOC as Data: A Thematic Approach to Caribbean Newspapers” grant application.

As the Scholarly lead and ex-Officio for dLOC’s Scholarly Advisory Board, I understand the need for accessible data and research from and about the Caribbean. This project’s advisory committee was selected due to their active participation and suggestions regarding the need to provide multi-lingual data-sets from the Caribbean to further scholarship and research. The project’s conceptualization emerged through conversations with members of our advisory committee about the lack of historical data on hurricanes and tropical storms in the Caribbean and how often generalizations have been created without evidence. We know that hurricanes and tropical storms have shaped the historical context in South Florida and the Caribbean, however historical hurricane data and the voices of Caribbean people are often not visible. This project seeks to amplify the narratives of these often marginalized groups.

Providing access to texts through dLOC’s extensive Caribbean historical newspapers will have far-reaching impact on scholars who study the role of disasters, land, climate change, colonialism, and other topics. We are looking forward to evaluating the quality of multilingual OCR and making suggestions for its improvement. dLOC is in a unique position to engage scholars, researchers, and teachers about the availability of these data-sets.

I look forward to working with my co-project lead, Perry Collins, dLOC Director Miguel Asencio, and Senior Administrative Lead Jamie Rogers. We firmly believe that this collaborative base, will be the first in future collaborations to make data-set available for scholarly research. I fully support the “dLOC as Data: A Thematic Approach to Caribbean Newspapers” project. Please let me know if you have questions.

Sincerely,

Hadassah St. Hubert, Ph.D.
CLIR Postdoctoral Fellow
Digital Library of the Caribbean (dLOC)