Odum Institute for Research in Social Science, University of North Carolina at Chapel Hill

Expanding the Dataverse: Developing capacity for advanced multilingual research data archiving

Project Justification

The Odum Institute for Research in Social Science (Odum; lead) in partnership with The Arab Council for the Social Sciences (ACSS; consultant) requests a $525,348 implementation grant under the National Leadership Grants for Libraries program for the development, testing, and integration of a new, fully internationalized front-end web interface for the Dataverse (DV) open-source research data repository software to enable handling of right-to-left (RTL) language scripts. **This project addresses chronic deficiencies in existing digital repository platforms to display RTL language content online, which is necessary for proper arrangement, description, and recruitment of content about or produced within various critically underserved language communities both in the US and abroad.** This unique effort draws on top talent and expertise to build the capacity of a widely used tool to address diverse global research community collaboration needs throughout the country and the world by increasing trust and ease of use through the addition of capacity for Arabic and other critical languages. This effort supports IMLS National Leadership Grants for Libraries Program Goal 3, Objectives 3.1, 3.2, and 3.3 by improving and expanding the ability of library and archives practitioners to manage and share research data from a diverse set of researchers and communities of interest. It continues existing collaborative work across multiple organizations to provide a long-overdue infrastructure upgrade to support user requirements for multilingual archival content.

DV is an open-source data repository platform that was originally developed by the Institute for Quantitative Social Science at Harvard. DV is developed primarily in the United States as an open-source software project by work contributed from many institutions around the country, including Odum. Now with 89 installations located across 6 continents, DV is the center of global efforts to preserve and share research data, with a growing international community of programmers, archivists, and researchers contributing to its ongoing development. This active community, which was recently formalized as the Global Dataverse Community Consortium (GDCC), has completed initial work to establish multilingual support for the DV. Due to limitations in the current web interface, this work has not yet included development of internationalization packages for RTL languages such as Arabic, though GDCC has noted demand for such.

This project is an effort to advance existing efforts at internationalization with four major contributions to data archival systems: 1) incorporation of a new DV interface mechanism that can handle RTL language scripts, 2) enhancement of general DV system capacity for RTL language scripts in both interface and metadata, and 3) creation of an Arabic internationalization package for DV, which will also serve as a prototype for packages in other RTL languages. Finally, 4) the Arabic-enabled DV will also serve as a model for other major digital repository platforms (e.g., Zenodo, DSpace, Fedora, Omeka, etc.) to implement similar solutions to serve a broader and more inclusive community of users. These contributions will benefit data archives serving a large community of researchers who collect data in Arabic and work with Arabic-speaking populations, with eventual benefit to those who work with speakers of other languages (e.g., Hebrew, Persian, Pushtu, Urdu, etc.), as well as the library and archives practitioners who support them.

Project Work Plan

Current literature identifies that Arabic-speaking researchers and those who work with Arabic-speaking communities are more likely to trust repositories and archives that use platforms that are built for the language in which they interact with participants and collect data. System and interface development for Arabic and other non-Latin languages tends to lag years, if not decades, behind English and other Euro-colonial languages. This is a particular problem for written languages that are not oriented left-to-right (LTR). ACSS in Beirut, Lebanon is a regional, independent, non-profit organization dedicated to strengthening social science research and knowledge production in the Arab world and the home of the first DV in MENA. ACSS will serve as a testbed and provide translation and testing support for this new development effort.

John D. Martin III (Research Data Systems Archivist at Odum), PI, works with information systems and online interfaces that handle or present information in Arabic and other RTL languages. He has a long record of experience working, studying, and conducting research with Arabic-speaking populations in the Middle East and North Africa (MENA). Dr. Charles Kurzman (Professor, UNC Sociology), Co-PI, has worked extensively to foster the development of international social science research partnerships with libraries and archives throughout MENA and previously worked with Odum to establish a DV for ACSS. Donald Sizemore (Odum) has served as the system administrator and primary maintainer of the ACSS DV since its establishment. Matthew Dunlap (Odum) has worked previously at Harvard IQSS as a developer on the DV project and continues to work with and make contributions to the project at Odum. Members of both GDCC and IQSS have expressed interest in and support for the efforts proposed for this project.

A web developer with Arabic-language experience will be tasked with designing, developing, and coordinating testing for the new RTL-enabled interface that will be deployed initially on the ACSS DV. A graduate assistant (GA) enrolled in the UNC School of Information and Library Science (SILS) will assist the PI and Co-PI with communication, information management, and reporting for the project and receive mentoring and training on project management, international collaboration, and research data archives practices and standards. The project staff will conduct program evaluation and will assess the project based on the following targets over the active period outlined below.

Year 1: targeted needs assessment, platform review, and code review (3 months); community discussions with both the development community and ACSS about goals and desiderata for development (3 months); beginning development phase; translation of internationalization packages with ACSS (6 months).

Year 2: active development of prototype interface in experimental fork of DV platform (6 months); iterative development and testing of prototype; iterative usability testing with ACSS (6 months).

Year 3: feature freeze and testing, bug fixes for prototype in experimental fork of DV platform (3 months); integration and testing through platform updates (3 months); negotiate integration of experimental fork with new interface into main DV platform fork (3-6 months); project closeout, reporting, and submission of any research testing and implementation findings for publication.

Diversity Plan

As discussed above, this project aims to increase access to research data management and archiving capabilities to a diverse, global community of library and archives practitioners who are currently underserved in this space. Through this development effort, researchers and their communities of focus will also be better represented through greater inclusion of contributed data in archives made publicly available throughout the world. The project staff will be intentional and mindful when recruiting new personnel and open-source software contributors to include members of underrepresented communities, particularly those in focus for this project. This effort will also deepen Odum's, UNC's, and the archives community's engagement with existing international partners and build bridges for new partnerships.

Project Results

The federal investment made through this grant will strengthen the ability of data librarians and archivists to serve a wider community of people. The product of this effort will address the chronic lack of Arabic-language and internationalization support for a US-led, open-source software platform which represents an intellectual and cultural export of this country. The project deliverables will be incorporated for use in future updates to DV such that as existing instances at institutions in the United States and abroad will receive the products of this effort as part of their normal upgrade cycles. The open-source nature of DV means that this work will have a lasting impact on all current and future users of the platform. This grant is an investment in the future of research data archiving and will benefit data archival practices as well as communities that are currently deeply underserved and underrepresented in the field.

Budget Summary

We request $525,348 from IMLS for the 3-year project, with UNC providing $532,062 in cost share for a total project budget of $1,057,410. Direct costs requested include $243,291 in salaries and wages with fringe benefits of $84,443 calculated at 26.753% plus pro rata fixed health insurance of $7,397; $20,713 in travel to conferences and project meetings to solicit community contributions and disseminate project outputs; $36,000 in ACSS contract fees; $2,500 in materials; and $48,105 for student support including 3 years of tuition ($14,052) and fees ($1,983). Indirect costs totaling $138,401 are calculated at the rate of 36% of modified total direct costs based on the F&A Rate Agreement. UNC's cost share includes $258,982 in salaries, $94,962 in fringe (calculated as above), $50,699 in tuition and fees, and $127,419 in indirect costs as calculated above.