Expanding the Dataverse:
Developing capacity for advanced multilingual research data archiving
NLG–Libraries–FY23 — IMLS Goal 3

1 Statement of need

The Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill (Odum, lead applicant) will work with The Arab Council for the Social Sciences (ACSS) and the Dataverse open-source community to address chronic deficiencies in existing digital research data repository platforms to display RTL (right-to-left) language content online, which is necessary for proper arrangement, description, and recruitment of content about or produced in various critically underserved language communities both in the United States and abroad. We request a $582,165 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 open-source research data repository software to enable handling of RTL language scripts. ACSS, located in Beirut, Lebanon, is the premier independent, non-profit social science research institute in the Middle East and North Africa (MENA) and home to the first public digital social science research data repository in the region. Odum has previously collaborated with ACSS in establishing this important regional archive using Dataverse, a leading open-source research data management platform. At the time ACSS Dataverse was established, the platform did not have the capacity to handle Arabic or any other RTL language in its interface, making the desired internationalization impossible at that time. Dataverse, as a platform, still does not support RTL languages. However, the Dataverse community is currently working on the early stages of a project to redevelop its front-end interface as a separate single-page application (SPA) that will have better extensibility and other contemporary features.

The *Expanding the Dataverse* project will align with this larger redevelopment effort to ensure that support for internationalization sufficient for the inclusion of Arabic and other RTL languages is integrated at the very earliest stages. This effort will address a need for supporting a wider range of critical languages in online digital data repository software. Using ACSS Dataverse as a prototype and test-bed, the proposed project will afford United States Arabic-speaking residents and scholars of MENA increased access to archival research data in a previously inaccessible language. Additionally, the project will provide archivists at ACSS and partner institutions like Odum enhanced ability to attract deposit of Arabic-language data from scholars in MENA countries. This unique effort draws on top talent with demonstrated expertise to build the capacity of this widely used tool to address diverse global research community collaboration needs in both the US and the world; by expanding the Dataverse to enable it to handle RTL languages, we will enable researchers throughout MENA to archive research materials in Arabic and other critical languages. This will increase trust and ease of use for researchers in underserved areas and increase United States researchers' access to research data that has until now remained un-archived, as has been the case with libraries and archives in MENA historically (Kurzman & Martin, 2018).

The *Expanding the Dataverse* project will serve several main audiences

1. **Librarians and archivists** who manage data relevant to populations speaking languages other than English/other Eurocolonial languages.

2. **Arabic-speakers** who currently do not have access to data archival platforms in their primary language.
3. **Social science researchers** whose work either produces or is aided by research data from Arabic-speaking countries.

1.1 The future of replicability on a global scale

Currently, Dataverse is the most-used open-source research data repository software in the world, serving research and institutional communities around the globe with 96 installations on six continents (see `https://dataverse.org`). It is a hugely important resource for research replicability and validation at a time when it is very much in the forward-thinking interests of researchers and their communities to publish and share their data for these purposes (Trisovic et al., 2020). The Dataverse project led by Dr. Gary King began in 2006 at the Harvard University Institute for Quantitative Social Science (IQSS) (King, 2007). It built on previous work from the Virtual Data Center (VDC) project and aimed at providing a platform through which research data could be archived and published for secondary use and to support the replicability, validation, and integrity of research (King, 2003). Development of the platform had its home in the United States at a time when internationalization of web-based software had not yet come to be a priority. Dataverse in its original state was only capable of providing an interface and handling information in the scripts associated with Eurocolonial languages: writing systems arranged from left-to-right (LTR). Now more than a decade and a half after it was first deployed, Dataverse is still unable to handle right-to-left (RTL) languages, nor does it have the capacity for proper internationalization. This is a critical gap, both in terms of equity and in terms of practical access.

Efforts to deploy Dataverse installations at institutions that serve populations whose languages use RTL or other complex scripts have been hampered by the inability of the platform to handle those scripts and languages. Web platforms and interfaces have been chronically slow to incorporate capacity for RTL languages. In the 2000s, when Dataverse and other digital repositories were becoming widespread, browser software was not capable of rendering RTL languages properly. However, in 2023, modern web software and large, well-funded proprietary web content platforms almost universally support RTL languages; in general, only independent and open-source platforms are lagging behind. But these platforms are the most likely to be accessible to Arabic-speaking countries or are most likely to be doing important work that needs to be available to lots of underserved people. It is particularly important that open-source platforms that host academic research be open and accessible to work in all types of languages.

1.2 Language equity in open-source

Part of the reason that open source platforms do not support RTL and the large commercially-supported platforms do is related to funding. Unlike commercially-supported platforms, which charge for their products or generate vast revenues from selling advertisement space, open-source platforms like Dataverse are usually not-for-profit and are developed by a large, disaggregated community of individual and institutional users and developers (Lerner & Tirole, 2005). Open source projects do not produce income to cover their development costs. Many rely on donations and grants for funding. Others, like RedHat and IBM open their source code, and rely on a professional services model to generate revenue to support software development.

To add capacity to an open-source project, institutions should fund the development time needed to add that capacity by allocating staff time to development. In this way, institutional users of platforms can

create solutions that meet their own needs and share those solutions with the larger project, improving it for everyone. Committing paid staff time allows institutions to help inform, influence, and steward the priorities and direction of open-source projects they use and contribute to. The Odum Institute seeks to do exactly this. As a public institution deploying open-source software to provide research data management solutions to a global community of users, the Odum Institute seeks funding in order to direct staff time to developing and supporting Dataverse, an open source software project that Odum and a worldwide community of users and institutions depend on.

1.2.1 Current and future Dataverse platform structure

Dataverse currently exists as a monolithic repository server platform. It relies on an articulated set of dependencies, including Payara application server, Solr search platform, PostgreSQL database engine, and others. Both the back-end and front-end of Dataverse are written currently in Java Enterprise Edition (EE) and are not separable from one another. When one part of Dataverse is deployed, then entire platform is deployed.

This means that the front-end (i.e. the interface that is shown to a user through a web browser) is lashed to the back-end (i.e. the platform server) and the two are inextricably linked. This model for development of web-applications is going out of vogue and many if not most no longer function in this manner.

Currently, web applications are developed using an application programming interface (API) that allows any arbitrary front-end interface to communicate with a server-side implementation of some system. Dataverse already has a robust web API that is avaiable, but the front-end of the platform currently does not implement all of the features available through the API, which leaves the front-end lagging, even though it is necessarily tied to the back-end.

One analogy might be to think of this as having a set of windows where there should be doors. We can see what is on the other side, but cannot access it directly because the wall is in the way.

The goal of the current Dataverse rearchitecture project is to put doors where those windows are by separating the front-end off as a second, entirely independent platform that communicates with the server-side back-end platform through the API. The goal of the *Expanding the Dataverse* project is to make sure those doors are able to open right-to-left (not to torture a metaphor) while also supporting the same as the monolithic Dataverse structure moves into its final stages of support and development over the next several years.

Odum staff has extensive experience in working with Dataverse both in development and system administration which positions us well to be able to aid in incorporating internationalization from the very beginning of the new front-end project. This is a unique opportunity and an inflection point for this platform and, by extension, for the world of data archiving to pursue a more equitable, inclusive agenda. This project will play a significant role in that pursuit.

1.3 Open source training and mentoring

In order for open source projects such as Dataverse to thrive, we need a workforce in archives, libraries, and other academic settings that understand how to work within open source communities, promote open source software as a viable alternative to vendor-based software solutions, collaborate inter-institutionally, and manage open-source development projects. Training for this set of skills is usually left to on-the-job experience, but this does not need to be the case.

As part of the *Expanding the Dataverse* project, we will recruit a Library and Information Science graduate student to the project and train them in project management and collaborative development. This serves two purposes: 1) the graduate trainee will gain experience in working with internationalization of an open source software platform, and 2) they will gain valuable and crucial project management skills that they would otherwise not have access to as part of their graduate program.

The downstream effects of having even one LIS graduate trainee finish their program with significant experience and ability for open source project management are hard to estimate. They will be able to bring these skills to bear in all subsequent work and also have the ability to train others in the same set of skills. These are also skills that are in very high demand, which will make better and more lucrative job opportunities available to them.

2 Impact

The *Expanding the Dataverse* represents a significant contribution to creating more accessible data archival systems and preparing a graduate trainee for engagement with both international, multilingual stakeholders, and open-source software development on a large scale. With three years of funding, these highly attainable targets will have tremendous impact in facilitating research data archives to reach a far wider audience in the long run and to attract important data deposits from critical language contexts.

2.0.1 Intended project results

The *Expanding the Dataverse* project will result in the following specific products:

1. A functioning multilingual (Arabic, English, and French) Dataverse interface hosted by ACSS.
2. Internationalization metadata packages for Dataverse.
3. Significant contributions to the Dataverse open-source codebase and sustainable engagement with community support for internationalization.
4. Successful progress toward program completion for a graduate trainee focused on data archives, UX, and open-source platforms.

The *Expanding the Dataverse* project aligns with 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. 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, with the immediate aim of making Dataverse accessible to Arabic-language research in the Middle East and North Africa. Arabic-speaking researchers and their communities of focus will be better represented in the global public archive through this contributed data.

This project will result in several important outcomes:

1. Develop capacity for research data archival systems to handle RTL languages in both interface and metadata schemata.
2. Establish internationalization user experience (UX) testing and evaluation protocols for data archival systems tailored to RTL languages to support platform accessibility.
3. Educate and mentor a graduate trainee in project management and open-source, collaborative development for data archival platforms.

4. Provide professional development opportunities for a graduate trainee through participation in professional meetings and on-site collaboration with international partners.
5. Continue to foster and expand existing international collaboration between US and MENA institutions such as Odum and ACSS.

3 Project personnel and resources

The personnel attached to this project provide its greatest strength. They are a dedicated, multiply-talented, highly-skilled team who have a great deal of previous experience working together. Below is a list of project personnel with a brief description of their major effort on the project, full descriptions are available in the *Project staff* document and in their *Resumes*.

John D. Martin III, PhD (Research Data Systems Archivist, Odum), PI, will guide and manage the proposed project, liaise with the Dataverse community, work with ACSS staff on translation packages, maintain the project codebase, and provide mentoring for the graduate trainee.

Charles Kurzman, PhD (Professor, UNC Sociology), Co-PI, will guide the proposed project and liaise with ACSS as a continuation of his previous work with the organization.

Thu-Mai Christian, MSIS (Assistant Director for Archives, Odum), Co-PI, will liaise with ACSS staff in developing user testing requirements.

Donald Sizemore, MSIS (System Administrator, Odum) will maintain development and testing environments for project staff, assist with code review, and guide the incorporation of project code into the main Dataverse codebase.

Akio Sone, PhD (Developer, Odum) will provide guidance and strategic planning for the development and expertise in integrating the work into the larger Dataverse codebase.

Victoria Hammett, PhD (Assistant Director for Education, Odum) will provide strategic planning, program evaluation, and mentoring over the course of the project.

TBN Web developer will be hired by the Odum Institute as a full-time permanent position to be the software development lead on the project.

TBN Graduate Research Assistant will be recruited from the UNC Chapel Hill School of Information and Library Science to provide project support and will be mentored in project management and open source, collaborative development.

Nada Chaya, MPH (Program Lead, Arab Social Science Monitor, Arab Council for Social Science) will coordinate translation and user testing services for the project.

For full budget information see the *Budget* and *Budget Justification*. We are requesting $582,165 from IMLS including indirect costs of $140,066, which includes $191,501 of student support. UNC is supplying 50.41% of the total cost of this project or $591,877. Major expenses for the project include:

Faculty/staff salary and benefits (70% paid as cost share).

Stipend, tuition, benefits, and fees for one graduate trainee for three years.

$45,000 for 6 hours per week of ACSS staff time over the three years of the project at a rate of $50/hour.

$2500 for a computer workstation for web development.

Travel support for project team members to attend an on-site meeting to work directly with members at ACSS in Beirut, Lebanon, attend and present at the Dataverse annual meetings for three years of the project, and attend and present at the RDA Plenary in the final year.

4 Project design

To achieve the goals listed above in the intended project results, we will divide the project into 5 primary components:

1. Platform assessment and beginning code review
2. Translation development
3. Feature development and testing
4. Code review, platform deployment, and testing
5. Project management and reporting

Work to be conducted by either Odum or ACSS staff or both will be indicated throughout the plan detail below. See the *Schedule of completion* (separate document) for a timeline related to the following activities.

4.1 Platform assessment and beginning code review

4.1.1 Needs assessment (ACSS)

ACSS will internally conduct a needs assessment based on their previous use of the ACSS Dataverse and the feedback of their stakeholder community. As part of the needs assessment, ACSS staff will recruit for and conduct two focus groups with researchers and other stakeholders currently using the ACSS Dataverse. Once the focus groups are completed, Odum and ACSS staff will collectively review the findings to guide development based on the needs of the community using the

4.1.2 Platform review (Odum)

Concurrent with the needs assessment, Odum staff will begin a platform review of both the existing Dataverse interface and the repositories containing code for the new Single Page Application interface to identify opportunities for including internationalization features in the existing code. As part of the platform review, Odum staff will conduct a comprehensive review and triage of issues related to localization, internationalization, and usability for Dataverse. The Odum staff will also conduct an analysis of other projects that have attempted to incorporate non-Latin scripts into Dataverse, such as the Fudan University Social Science Data Repository.

4.2 Translation development

4.2.1 Metadata key translation (ACSS)

ACSS staff will review and translate the metadata key packages for Dataverse used for localization and internationalization. These packages contain lists of keys linked to either functional or conceptual terms that appear throughout the interface. For instance, the text that appears on buttons or field labels in the interface. Each of these 200+ keys has to be individually translated and then verified as aligning or matching conceptually with those in other languages. Odum staff will review these with ACSS staff and create pull requests to get them into the localization and internationalization packages. The translated keys will be reviewed again at several stages throughout the project to ensure that they remain current with the state of the interface code as it develops.

4.2.2 Documentation translation (ACSS)

In addition to the metadata keys, ACSS staff will translate and update the Dataverse guides user and developer documentation for the platform and the API. This will ensure that Arabic-speaking developers and system administrators will be able to operate on the platform going forward and will hopefully attract interest in its future development. Documentation translation will also be regularly reviewed throughout by both ACSS and Odum staff to ensure that it remains current and up-to-date.

4.3 Feature development and testing

4.3.1 Source code forks (Odum)

Odum staff will create and maintain development forks (parallel versions of existing code) from the Dataverse codebase in GitHub for both Dataverse core code and the new Dataverse front-end code. Development will be continuous and rapid throughout the project to keep apace with the rest of the Dataverse community.

4.3.2 Usability development (Odum)

Odum staff will focus on usability in the interface to ensure that information flow and visual presentation are congruous with the expectations of users for how the interface should respond. This will include both basic targets such as the placement of buttons and links and other interface elements on screen as well as more advanced features such as ensuring the appropriate cascading of interface elements on different screen shapes and sizes.

4.3.3 Localization and Internationalization development (Odum)

Odum staff will incorporate the translation packages developed by ACSS staff into the design and implementation of the interface. Primary development will occur with Arabic in mind, but other languages will also come into play, such as French, which is also commonly spoken and read throughout many MENA countries. The goal will be to have a seamless user experience across languages and, in some cases, seamless switching between them.

4.3.4 Iterative testing (ACSS)

Throughout the development process, ACSS staff will provide user testing support. Odum staff will maintain separate development and testing servers so that we will not be operating directly on the production environment for ACSS at any point. Whenever Odum staff make an update, ACSS staff will test and work with the updated interface to ensure that it is working as expected. In cases when it does not, ACSS staff will create issues in the Odum GitHub repositories associated with our Dataverse source code forks. These are like service tickets and will allow Odum staff to track and address bugs and other unexpected behaviors.

4.4 Code review, platform deployment, and testing

4.4.1 Prototype testing (Odum and ACSS

Prior to the actual deployment of the new ACSS Dataverse front-end interface, Odum staff will stand up a fully-functional clone of the existing ACSS Dataverse containing all of the same data, but with a new interface on an updated platform. This prototype will be rigorously and extensively tested by ACSS and Odum staff prior to final deployment.

4.4.2 User testing (ACSS)

ACSS staff will again recruit users from their community to test the prototype platform. This time, instead of focus groups, user testing will occur in vivo with followup questionnaires for user-testers.

4.4.3 Code review and bug fixing (Odum and ACSS)

Once the prototype ACSS Dataverse has completed user testing, Odum and ACSS staff will review the findings of the testing and make changes to address bugs in the interface. This will ensure that major issues do not make it into the production version of the ACSS Dataverse when it is finally deployed.

4.4.4 Production deployment (Odum)

After the prototype ACSS Dataverse has been sufficiently debugged, Odum staff will work with the core Dataverse team to incorporate changes into the codebase that have not already been integrated. Once code changes have been integrated, we will begin the process of upgrading the production instance of the ACSS Dataverse to the new platform and interface for final deployment. By the end of this process, the ACSS Dataverse will be running a fully updated, fully internationalized platform at the cutting edge of data archival repository technology. Additionally, the Dataverse project will have incorporated a more widely-accessible set of language capacities into its codebase.

4.5 Project management and reporting

4.5.1 Annual reporting (Odum)

The PI and Co-PIs will produce and distribute an annual report on project progress and expenditures for the year prior. This will be done in consultation with the rest of project staff but will primarily be the responsibility of the PI.

4.5.2 Dataverse annual meeting (Odum)

In conjunction with the annual report, project staff will plan to submit an abstract for presentation at the Dataverse Annual Meeting, typically held in Boston, in June of each year. The PI and one Co-PI will attend the meeting along with the web developer and GRA on the project. This will serve as an opportunity to meet and work directly with members of both the Dataverse core development team and the extended Dataverse community to discuss the work being conducted as part of the *Expanding the Dataverse* project. It will also provide an opportunity for professional development for the GRA within and open source community.

4.5.3 Quarterly full-team meetings (Odum and ACSS)

Each quarter, we will schedule a one-hour Zoom call with an agenda to check in on the high-level goals and timeline and ensure that the project is moving forward according to the intended schedule. This will also be an opportunity to discuss new ideas that have arisen as part of the larger effort at hand and incorporate them into planning for the next quarter.

4.5.4 Weekly project check-in (Odum and ACSS)

Every week we will schedule a one-hour Zoom call with staff to address any concerns or issues that arise as part of the day-to-day operation of the project. This may not always include the entire staff, but will include a more targeted set of staff depending on which phase of the project we are currently in.

4.5.5 ACSS project on-site meeting (Odum and ACSS

During the second year of the project, the PI, one of the Co-PIs, and the GRA will travel to Beirut, Lebanon for 5 days to meet on-site at ACSS and work together to develop the ongoing user testing targets to facilitate iterative testing, review, and feedback as development proceeds.

4.5.6 ACSS on-site extended work (ACSS and Odum GRA)

After the on-site meeting at ACSS in Beirut, Lebanon, the project GRA will remain for two weeks to continue to work directly with ACSS staff as the user testing activity begins. The goal of this activity is for the GRA to gain experience working with an international partner and to develop more autonomy as a project manager.

4.5.7 Research Data Alliance Plenary (Odum and ACSS)

In the final year of the project, several members of the Odum project staff and one member of ACSS staff will attend the Research Data Alliance (RDA) Plenary meeting where they will give a talk about the project and offer a demonstration. This will also afford project staff the opportunity for professional development and to engage with top experts in the field of research data management and curation.

5 Project results

This project will have a lasting impact on the accessibility and usability of digital data archival repositories around the world. Through federal support of the *Expanding the Dataverse* project, IMLS will ensure that more people are able to access data in more languages than ever before through the most widely-used digital data repository platform currently available. 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 Odum Institute is one of the largest institutional users of the Dataverse project software. In addition to deploying, administering, and maintaining multiple installations of Dataverse for UNC and other institutions, Odum is home to developers who have ties to the community and work directly on aspects of the project or parallel projects that add features to the software. Odum is also an institutional home for the Global Dataverse Community Consortium and maintains a global set of institutional relationships related to that effort. The institute is very much a part of the Dataverse community, make it an excellent home for an effort such as the *Expanding the Dataverse* project to succeed.

Additionally, Odum's history of partnership with the Arab Council for Social Science allows us to create and enhance a resource that benefits not only members of that institution's constituent community overseas, but the many social science researchers in the United States whose work is aided by having access to data deposited in the ACSS Dataverse.

The open-source nature of the Dataverse platform means that this work will have a lasting impact on all current and future users of the platform. Contributions from this project to the Dataverse codebase will be incorporated for use in future updates to Dataverse such that existing instances at institutions in the United States and abroad will receive the products of this effort as part of their normal upgrade cycles. 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.

References

King, G. (2003, March). The future of replication. *International studies perspectives*, *4*(1), 72–107. doi: 10.1111/1528-3577.04105

King, G. (2007, November). An introduction to the dataverse network as an infrastructure for data sharing. *Sociological methods & research*, *36*(2), 173–199. doi: 10.1177/0049124107306660

Kurzman, C., & Martin, J. D. (2018, December). The hidden heritage of arab libraries: Online catalogs and institutional barriers to discoverability. *IFLA Journal*, *44*(4), 300–310. doi: 10.1177/0340035218785189

Lerner, J., & Tirole, J. (2005, June). The economics of technology sharing: Open source and beyond. *The journal of economic perspectives: a journal of the American Economic Association*, *19*(2), 99–120. Retrieved from `https://www.aeaweb.org/articles?id=10.1257/0895330054048678` doi: 10.1257/0895330054048678

Trisovic, A., Durbin, P., Schlatter, T., Durand, G., Barbosa, S., Brooke, D., & Crosas, M. (2020, June). Advancing computational reproducibility in the dataverse data repository platform. In *Proceedings of the 3rd international workshop on practical reproducible evaluation of computer systems* (pp. 15–20). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3391800.3398173

See *Narrative* for detailed project design descriptions of below tasks.

Year 1

Q1

Needs assessment ACSS staff (6 months)
Platform review Odum staff (6 months)
Recruit and hire web developer Odum staff (1–2 months)
Recruit GRA Odum staff (1 month)

Q2

Dataverse GitHub issue review PI / web dev (6–9 months)
Recruit focus groups ACSS staff (1 month)
Conduct focus groups ACSS staff (3 months)

Q3

Metadata key translation ACSS staff (6 months)
Review focus group findings Odum and staff (3 months)

Q4

Attend Dataverse Annual Meeting Odum staff
Translation verification Odum and ACSS staff (1 month)
Prepare translation packages Odum staff (6 months)
Begin initial development code forks Odum staff / web dev (3 months, then ongoing for 12 months)
Annual report PI (1 month)

Year 2

Q1

Translation review ACSS staff and PI (3 months)
Documentation translation ACSS staff (9 months)
Begin initial iterative testing ACSS staff (6 months)
Basic usability development Odum staff / web dev (6 months)
Basic L10N/i18n development Odum staff / web dev (6 months)

Q2

Recruit community user testers ACSS staff (3 months)
ACSS Beirut, Lebanon on-site meeting ACSS staff / PI / Co-PI / GRA (5 days)
ACSS Beirut, Lebanon extended on-site user testing ACSS staff / GRA (5 days)

Q3

Initial community user testing ACSS staff / community (3 months)
Advanced usability development Odum staff / web dev (6 months)
Advanced L10N/i18n development Odum staff / web dev (6 months)
Issues in DV core and ACSS fork Odum staff (18 months)

Q4

Attend Dataverse Annual Meeting Odum staff
Code review Odum staff (3 months)
Prototype deploy Odum staff (3 months)
Pull requests to DV core Odum staff (15 months)
Annual report PI (1 month)

Year 3

Q1

Targeted bug smashing Odum staff (6 months)
Translation review ACSS and Odum staff (3 months)
User testing review Odum staff 3 months

Q2

Pre-deploy production testing (beta) Odum and ACSS staff ()
RDA meeting attend Odum and ACSS staff

Q3

Production deploy Odum staff (3 months)
Post-deploy production testing ACSS staff (3 months)

Q4

Final meta translation review Odum and ACSS staff (3 months)
Final documentation translation review Odum and ACSS staff (3 months)
Final documentation compile Odum staff (3 months)
Annual report PI (1 month)

Section I: Intellectual property rights and permissions

Digital products description

The digital products from this project will take two forms:

1. Software source code related to The Dataverse Project, an open source digital research data repository platform. All source code will be housed in repositories on GitHub managed by the Odum Institute at the University of North Carolina at Chapel Hill. Ultimately, all reviewed and tested code will be contributed in the form of pull requests to the main fork of the Dataverse repository.
2. Reports, white papers, articles, presentation materials, evaluation rubrics, and other documents related to the project and its activities. All such documents will be made publicly available in the form of plain text or PDF file formats, or both.

Licensing and digital ownership rights

The project team will license all source code produced under an Apache 2.0 license, which the license under which all Dataverse Project source code is distributed. The terms of the license allow for use of the software for any purpose, including reproduction, distribution, modification, and the production of derivative works without any further permissions, royalties, or restrictions except for a requirement that the following be included in any downstream code:

1. The original copyright notice;
2. A copy of the Apache 2.0 license itself;
3. If applicable, a statement of any significant changes made to the original code;
4. A copy of the NOTICE file with attribution notes (if the original library has one).

Section III: Software

General information

Description of software created

The project will develop, test, and integrate 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.

Similar existing software

The PI is not aware of any existing data repository platforms that offer a front-end web interface that effectively accommodates RTL-languages. The RTL language enhanced Dataverse will build the capacity of a widely used data repository 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.

Technical information

The code will be primarily made as contributions to the Dataverse project code and will therefore conform to the languages used in the project itself. Currently the core of the project uses Java Enterprise Edition (EE). The new SPA front-end application is React written in JavaScript and TypeScript and runs on NodeJS.

Programming languages

The code will be primarily made as contributions to the Dataverse project code and will therefore conform to the languages used in the project itself. Currently the core of the project uses Java Enterprise Edition (EE). The new SPA front-end application is React written in JavaScript and TypeScript and runs on NodeJS.

Interoperability

The code will largely exist in the core of the Dataverse project code as it will be subsumed in the form of pull requests to extend the international language capabilities of the Dataverse repository software platform.

Dependencies

No underlying additional software or system dependencies will be necessary to run the enhanced Dataverse software. Only those dependencies required for Dataverse will be needed.

Documentation

Dataverse has community standards for documentation in place. This project will follow those standards in contributing code and documentation, represented by the documentation listed at: https://guides.dataverse.org.

Previous software repositories

Dataverse Toolbox https://github.com/OdumInstitute/dataverse-toolbox
Dataverse CoReRe https://github.com/OdumInstitute/dataverse-corere
Dataverse Ansible https://github.com/OdumInstitute/dataverse-ansible
Dataverse Aux https://github.com/OdumInstitute/dataverse-aux
Dataverse Automated User Tests https://github.com/OdumInstitute/dataverse-automated-user-tests
R Binary File Reader https://github.com/OdumInstitute/R-binary-file-reader
TRSA Web https://github.com/OdumInstitute/trsa-web

Access and use

Availability

The code for expanding the Dataverse with RTL capabilities will remain open source. Code and all other digital products and documentation will be maintained in a public GitHub repository and made available to software developers and users.

Source code repository

The code will be available in the main Dataverse core code repository and the new Dataverse front-end repository upon project completion.
Dataverse https://github.com/IQSS/dataverse
Dataverse Frontend https://github.com/IQSS/dataverse-frontend
It will be incorporated into existing project code and will not exist as a standalone repository.

H. W. Odum Institute for Research in Social Science

Founded by Howard W. Odum in 1924, the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill (UNC) is considered the oldest interdisciplinary social science institute at a research university in the United States. Although almost a century has passed since its inception, the Odum Institute remains steadfast in its mission to foster groundbreaking social research that improves the lives of people in North Carolina and around the world. Consistent with this mission, the Odum Institute Data Archive has worked steadily to keep pace with researchers' data management, preservation, and sharing needs, making the Institute a national leader in the development and implementation of tools and services for research data management and preservation.

The Odum Institute Data Archive was formally established in 1969 when it received funds from the National Science Foundation to create an academic center of excellence in science to include computing facilities for a social science statistical laboratory and data center. Today, the Data Archive houses one of the oldest and largest catalogs of machine-readable data in the United States. The Odum Institute is a founding member of the Data Preservation Alliance for the Social Sciences (Data-PASS), which is a voluntary partnership of the largest and most influential social science archives in the U.S. that works to establish and promote sustainable, evidence-based archival standards, technological solutions, and business models for preserving "at-risk" social science data. The Archive is also represented on the board of the International Federation of Data Organizations (IFDO), which coordinates international efforts to meet the increasingly complex needs of the social science research community worldwide.

The Odum Institute has adopted Dataverse as its virtual archives platform, which allows scholars from around the world to store, manage, analyze, and share their data using its user-friendly web interface. The UNC Dataverse hosted by the Odum Institute Data Archive provides access to over 25,000 datasets and 235,000 files including the exclusive Louis Harris Data Center collection, the National Network of State Polls, and the most complete holdings of 1970 Census data files. In response to growing stakeholder demand for research transparency and data access from the scientific community, the Odum Institute Data Archive has established a suite of data management and curation services that provides support to scholars throughout the research lifecycle including data management planning, curation, training, and archiving. The Data Archive also offers a unique replication data verification service to academic journals seeking assistance with data policy implementation and enforcement.

As an early adopter of the open source Dataverse software, the Odum Institute has been actively engaged in ongoing Dataverse development, making critical contributions to the Dataverse codebase, which is maintained by the Institute of Quantitative Social Science at Harvard University. This experience has enabled the Odum Institute to provide Dataverse installation and hosting services to other institutions that require expertise and technical infrastructure beyond what their existing resources can support. The Odum Institute also leads the Global Dataverse Community Consortium (GDCC), which provides organization to the international community of organizations who have adopted Dataverse as their repository platform. GDCC also serves as a venue for collaboration among these organizations to leverage economies of scale in support of Dataverse repositories around the world.

The Odum Institute Data Archive designated community consists of members of the academic community at UNC who are engaged in social research or research in other disciplines that includes a social component. These members include research faculty, students, and other individuals that participate in such academic research. Because the Odum Institute Data Archive provides free and open access to its collections, data are also accessed and used by journalists, policymakers, citizen scientists, and others around the world who are interested in the study and understanding of society and social relationships.