2022

Library-Based Data Curation, Management and Interdisciplinary Research at Florida International University: Reciprocal Use of Data through Collaboration

Zhaohui Fu
Florida International University, Fujen@fiu.edu

Levente Juhasz
Florida International University, ljuhasz@fiu.edu

Jill V. Krefft
Florida International University, jkrefft@fiu.edu

Boyuan Guan
Florida International University

Library-Based Data Curation, Management, and Interdisciplinary Research at Florida International University: Reciprocal Use of Data through Collaboration

Abstract

This paper shows the Florida International University (FIU) Libraries' efforts in research data management, implementation, and practice. The FIU Library-based data team has led research projects that are directly built upon institutional and other data repositories, data hubs, and data visualization tools, through collaboration with the user community. We will present the technology setup and configuration of such a data framework, which includes Dataverse, ESRI's ArcGIS Data Hub, and other data collection and visualization tools. We will also discuss the fiscal and organizational structure needed to support research data management initiatives. Using a couple of our applied research projects, we will demonstrate how users interact, collaborate, and contribute. We will discuss the challenges we encounter when it comes to data sharing, data curation, management, and most importantly, serving users at all levels of data literacy. In addition, we will present from a researcher's perspective how he/she can manage his/her publications and associated research data and make them discoverable. We will showcase the websites of interdisciplinary research projects, such as the FIU Jack D. Gordon Institute for Public Policy's (JGI) Security Research Hub (see also: https://srh.fiu.edu/home/) and the USAID-funded Global Water for Sustainability Program (see also: http://dpanther.fiu.edu/glows/). These project websites use FIU institutional repository and research data management platforms (see also: http://rdm.fiu.edu/dataverse/). We will demonstrate how we support a community-based research project using FIU's research data repository and data hub, as well as community-based data repositories such as ESRI's Living Atlas of the World.

Introduction

The concept of library-based research data management and curation is not a novel one. The Association of Research Libraries (ARL) SPEC Kit 354 provides a detailed study on the current state of research data curation and management among US academic libraries (ARL, 2017). Academic libraries establish data management services and digital data stewardship mainly to serve faculty and students and are largely influenced by the federal funding agencies' mandate for research data sharing. Among all the data platforms and other digital repository platforms used, the Harvard-developed open-source platform "Dataverse" proves to be the most relevant and mature academic research data management tool. Different from other public domain data repositories, Dataverse uses DOIs and Primary ID to support proper attribution to the data creator and support long-term preservation in line with FAIR principles. Many academic libraries have been utilizing Dataverse for their research data management purposes, either contributing directly to Harvard's Dataverse or through local installations of the Dataverse software. The Florida International University Libraries have created their own Dataverse incidence in collaboration with the' Department of IT (DoIT) at Florida International University (FIU), and

through the support of FIU Technology Fee Fund (tech fee) grants, to establish a mature and sustainable research data repository system. Tech fee grants are proposed by students, faculty and staff and are selected based on a competitive process each year.

While these academic platforms have unified the practice of data management and preservation, few library-based research data management initiatives have ventured into the arena of participatory data use and community-based research, where open data can be connected and integrated through data management and visualization tools. In the universe of data curation, data repositories, and data management, academic libraries are relatively new and operate on a small scale, due to motivation, fiscal limitation, administrative structures, human resources and expertise, as well as capacities in IT framework and networking (Yoon, 2019). Domain-specific data repositories, as well as government data warehouses and public domain data collections, remain as the main data access points for most users, including academic ones. The Knowledge Network for Biocomplexity (KNB) is one example of a standardized, mature data repository that fully meet the needs of the scientists within its domain of specialization. KNB is built upon Data Observation Network for Earth (DataONE), an NSF-funded initiative for providing a distributed framework and sustainable cyberinfrastructure that meets the needs of science and society for open, persistent, robust, and secure access to well-described and easily discovered Earth observational data. A domain-specific data repository as such is largely utilized by scientists and researchers for data sharing and fulfilling their obligations to federal funding agencies.

Another dimension of data management and curation efforts emerges from a blending of commercial vendors with government and community participants. The Environmental Systems Research Institute's (ESRI) ArcGIS Data Hub is a good example. As a main international supplier of geographic information system (GIS) software, ESRI, over the years, has attracted many government agencies and research institutions all over the world to their platform, to become a dominant player in geospatial data software. Its newly established ArcGIS Data Hub has quickly become a major cloud-based geospatial data sharing platform.

Academic data librarians and data specialists have quickly realized that to serve faculty and students in terms of data curation and management, they often have to navigate among all these different dimensions within the data universe. This requires making connections and establishing gateways among these dimensions, increasing data literacy among students and researchers. By innovating and applying existing resources, tools, and platforms, we can grow the user base, encourage participation, and, ideally, foster direct contribution from our researchers and students. (See Table 1)

Dataverse	DataONE	ArcGIS Data Hub (ESRI)
Academic Libraries	Domain Research Network	Government, Community, and Vendors
Harvard	DataONE	ESRI
Strong focus on Arts and Humanities; Social Sciences	Earth and Environment	Geospatial

DOI, Primary IDs, FAIR Principles Standardized Metadata (DDI Lite, DDI 2.5, Dublin Core); Controlled vocabulary (ISO 63901, ISO 3166-1, OBI Ontology and NCBI Taxonomy)	DOI, Primary IDs, FAIR Compliant, Standardized Metadata (Dublin Core, EML, ISO 19137, GDC CSDGM, ISO 19115, DDI); Controlled vocabulary (Domain specific Ontology and Taxonomy)	Interoperable with ArcGIS Online data visualization and analysis tools; Standardized metadata, (GDC CSDGM, ISO 19139 GML3.2, ISO 19139 ISO 19115 2003) Controlled vocabulary for geographic names

Table 1: Comparison of Dataverse, DataONE, and ESRI ArcGIS Data Hub

This article serves as a case study of Florida International University Libraries' data curation and management, how data librarians and specialized library faculty play leadership and facilitation roles in data-driven interdisciplinary research, and how we engage students, researchers, and the community in participation.

FIU's Research Data Repository

The FIU Research Data Portal was established in 2017, after many years of planning and building partnerships across campus to gain support from various academic departments. The FIU Research Data Portal serves as an open access interdisciplinary data archive, supporting access and reuse of unique institutional research data produced at FIU. Members across departments and disciplines at FIU, including Biological Sciences, History, Digital Humanities, Computer Science, Engineering, and Health Sciences all utilize the data repository to share and preserve their data.

The Research Data Portal is built on the Harvard Dataverse open-source web application, to help faculty and students share, preserve, cite, explore, and analyze research data. The Research Data Portal was funded through two institutional tech fee grants totaling $487,241. Most of this funding was utilized for IT cloud storage, online services, and subscriptions, as well as equipment. The FIU Library Digital Collections Center and the FIU Library Geographic Information Systems (GIS) Center collaborated with the FIU Division of IT, as well as the FIU Office of Research and Economic Development for sponsored research, in the planning and implementation process. FIU Library administration redefined the job descriptions for two existing positions to support research data curation and management, while redefining research associate positions within the GIS Center to include geospatial data and metadata services. The team also includes programmers and IT support staff from both the GIS Center and the Library Systems department. The Institutional Repository coordinator led the implementation effort in data curation and management, as data curation falls into the category of preservation and sharing of institutional intellectual property, and publications and data often go hand in hand.

To achieve fiscal efficiency, our approach was to begin the planning process using existing human resources, to gain expertise by encouraging motivated staff to further their professional development and training, and to innovate by writing internal tech fee proposals to gain fiscal support. The team succeeded in securing two tech fee grants, which helped to establish the IT infrastructure needed for the Research Data Repository.

The two tech fee grants provided ample storage for smaller datasets, less than 10GB, from students and researchers for the first five years of the program. As the program matures and grows, additional funding sources will be necessary. We will continue to seek internal funding but will also leverage

partnerships with research groups across campus. This will include providing fee-based services for larger amounts of data, which can be written into grants.

Built upon the Harvard-developed Dataverse, FIU created its own Dataverse incidence. This mainstream open-source platform streamlines workflows, allowing authors to easily upload data and create metadata. Features within the system include integration with DataCite to assign DOIs, versioning, generation of UNF (universal numerical fingerprint) for tabular data, API, and metrics via Make Data Count.

Each dataset contains three levels of metadata: citation metadata, domain-specific metadata, and file-level metadata. Dataverse employs standards-compliant metadata to ensure interoperability. Citation and domain-specific metadata schemas supported within Dataverse include DDI Lite, DDI 2.5 Codebook, and Dublin Core. In addition, Dataverse incorporates several discipline taxonomies, including IS0 639-1 for Languages, IS0 3166-1 for Country/Nation fields, OBI Ontology, and NCBI Taxonomy for Organisms. (Dataverse Metadata References). Discipline is important because richer metadata supports greater access and reuse. All metadata can be exported into JSON, Dublin Core, DDI HTML Codebook, OAI_ORE, and OpenAIRE, allowing metadata to be easily ingested into various systems without the need to crosswalk or transform metadata.

The Dataverse-based research data management platform plays an important role in the overall data services that the FIU Libraries provide for their students and faculty. The RDM team in the library provides consultation to students and faculty relating to metadata creation, data management, copyright, and access. While the users typically create metadata and upload their data packages, the team performs quality control.

FIU's Geospatial Data Framework and Data Services

The FIU Library GIS Center has maintained an enterprise geodatabase holding geospatial datasets created by the GIS Center and affiliated researchers since 2005. We created the FIU Geoportal utilizing a combination of KNB's EML metadata schema and editor to create metadata, and ESRI's Geoportal to publish metadata and serve as the user interface (UI). FIU developed a comprehensive metadata creation procedure, which was used to create metadata records for each dataset. Over time, the FIU Geoportal has hosted 1,273 geospatial datasets, some of which were created by FIU researchers through grant activities. In 2020 and 2021, we began the migration of our data to an off-the-shelf data catalog and discovery solution from ESRI, ArcGIS Hub Premium, to provide access to geospatial datasets we curate or create. ArcGIS Data Hub is an extension of ESRI's ArcGIS Online platform, and organizes items based on item-level tags in the metadata record, which allows for grouping and discoverability of datasets. An item-level tag is required for any item in ArcGIS Online, while other components of the metadata record, such as a summary and description, how accurate and recent the item is, and restrictions associated with using and sharing the item, may or may not be available. Various metadata styles are available within ArcGIS Online and ArcGIS Data Hub, but the default and most common is the North America Profile of ISO 19115 2003 (Metadata in the ArcGIS Online Ecosystem, 2022). Other metadata styles available within ArcGIS Hub are: FGDC CSDGM Metadata, INSPIRE Metadata Directive, ISO 19139 Metadata Implementation Specification GML3.2, and ISO 19139 Metadata Implementation Specification (Metadata in the ArcGIS Online Ecosystem, 2022). Datasets generated or curated by FIU researchers may be uploaded to ArcGIS Online and shared with the community via ArcGIS Hub. When an item/dataset is uploaded to ArcGIS Online, a metadata record is created automatically, with item-level

tags required and driving the metadata record. ArcGIS Data Hub can be used in concert with Dataverse, as each item created has a unique URL. In cases when a unique DOI is needed, Dataverse must be used, because ArcGIS Data Hub does not generate these unique IDs.

As we piloted ArcGIS Hub for storing and sharing our original and curated geospatial data, we adapted our already robust metadata creation procedures to suit the updated guidelines within the ESRI ecosystem, and deployed ArcGIS Data Hub Premium for a specific project, the FIU Security Research Hub. ArcGIS Data Hub Premium allows us to collect and curate datasets created by FIU researchers and members of the community, as well as providing access to hundreds of authoritative datasets via ESRI's Living Atlas, some of which are publicly available, while others are available by subscription only (ESRI's Living Atlas of the World, 2022). The FIU Security Research Hub (SRH) data collection operationalized with ArcGIS Data Hub serves as the pilot project, and we plan to begin the migration of our other datasets in the near future.

Data services at FIU include online resources in the form of LibGuides related to data literacy, data sources, and tools. However, the Library offers services that go beyond the traditional reference guides. The Libraries' Digital Collections Center and Geographic Information Systems Center, through their GIS labs, Digital Scholar Studio, and Institutional Repository, as well as research data management venues, provide data preparation, processing, analysis, visualization, text-mining, and story-telling services. We have established a dPanther IT computing platform, combining local and cloud computing environments, to enable students and faculty to perform the above-mentioned data functions. We routinely provide synchronous and asynchronous introductory workshops on data and tools.

The Security Research Hub Project -- An Effort Led by Library Data Professionals

Florida International University's (FIU) Security Research Hub is a virtual research platform that aims to harness publicly available information which supports collaboration and shared understanding about Latin American and Caribbean security issues amongst FIU and its partners. The Hub will focus on a wide range of security-related topics such as transnational organized crime, migration, illegal fishing, and regional health concerns.

Within the SRH, FIU provides a suite of analytic, geospatial, and other research tools that allow researchers to process and analyze large amounts of data harvested from open-source, publicly available information, and other licensed data/subscriptions. The SRH aims to solicit input from partners, and it draws from FIU's expertise in its schools, centers, and preeminent programs. These include the Institute for Public Policy, the Library Geographic Information Systems (GIS) Center and Digital Collections Center, the Latin American and Caribbean Center, the Global Forensics and Justice Center, the Global Health Consortium, and the Extreme Events Institute.

The core technical team for the SRH is largely composed of library faculty and researchers from the GIS Center and the Digital Collections Center. One of the main challenges in creating such a research hub is how best to aggregate existing data and resources (including geospatial data, reports, publications, news articles, and web resources) on the above-mentioned security topics and serve as a one-stop shop for security research. The solution lies in exploiting what library professionals do best – managing large information networks such as the Digital Commons Institutional Repositories and utilizing existing community data repositories such as ArcGIS Data Hub and Dataverse. Standardized metadata

harvesting, editing, and creation are essential to the SRH. Geoweb visualization, data analysis, and mining tools are also included to facilitate discovery and use of the data and information.

There are four main components that the SRH offers: 1) News and Events; 2) Publications and Reports; 3) Data; 4) Geo-Dashboard and Infographics on any of the given security topics. One example of a user research case would be the most recent earthquake in Haiti and subsequent increased security threats (e.g., kidnapping). The SRH first filters all relevant news and social media information on this topic, then offers the reports and publications curated on the topic through the Digital Commons platform. If the user would like to use data and visualization, he/she can then navigate to ArcGIS Data Hub to explore further options on the topic. Within the SRH, we have utilized the Bepress API to ensure that all content (reports, publications, web resources) curated by the project team are aligned with the security themes and topics pre-selected by the project team. The API then interacts with our SRH UI to display only relevant entries when filters are applied.

Both Digital Commons and ArcGIS Data Hub also allow authenticated users to contribute back to the SRH. Within ArcGIS Online environments, we have used tools such as Survey123 and StoryMaps to solicit crime or kidnapping incidences spotted in the media or other venues. In areas where data is severely unstructured, unshared, or missing, we have also applied text-mining efforts to draw data points. For instance, our team went through articles on Haitian kidnapping and extracted times, locations, and persons involved in kidnapping incidences, mapped these incidences using Survey 123, and told a visual story using StoryMaps. These StoryMaps can be kept alive as new kidnapping incidences are collected through community sourcing using Survey123. Through the SRH project, the library's data curation and management effort is no longer a passive one--waiting for researchers to contribute data to the data catalog--but a proactive one, leading the researchers and community to both use and participate.

Utilization of Digital Commons and ArcGIS Data Hub academic subscriptions, as well as existing Dataverse platforms to support research projects, is a cost-effective way to benefit sponsored research in a profound way. Library data and information professionals can contribute directly to the funded research activities and interact with domain experts to gain understanding of data usage and usability.

Researchers' Experiences and Perspectives

While working with faculty on archiving and sharing data, we have learned from a few researchers that they feel more comfortable submitting their unique data to the institutional repository rather than sharing through other repositories such as GitHub or publisher-based repositories. The primary reason expressed by faculty is that they trust the institutional data repository over third-party repositories. This observation at an institutional level requires further exploration and is a possible area of research that could inform institutions when considering an institutional data repository. There is currently limited literature related to faculty data archiving and repository preferences. The current literature focuses more on whether faculty are likely to share their data (Piwowar, 2011) and differences in sharing data across disciplines (Akers, 2013).

Some of the main job components of university faculty are research, teaching, and service, with research being one of the primary foci for many faculty at universities classified as "R1: Doctoral Universities – Very high research activity" in the Carnegie Classification of Institutions of Higher Education. Florida International University is one example. Although there are variations in how research productivity is evaluated among institutions and job roles, generally the quality and quantity of

publications matter. This puts pressure on faculty to publish frequently, which is often described by the phrase "publish or perish" (see e.g., De Rond & Miller, 2005). At the same time, research reproducibility as a major scientific concept is also gaining momentum. According to the US National Science Foundation (NSF), "reproducibility refers to the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results". Reproducibility is a minimum necessary condition for a finding to be believable and informative." (Cacioppo et. al, 2015). In practice, however, reproducibility is not adapted smoothly by everyone, and there is an ongoing reproducibility crisis due to the lack of consensus on what reproducibility should be, and how exactly it should be done (Baker, 2016). Nevertheless, there is a noticeable shift in publishing practices in recent years, where journals and their publishers require or strongly encourage authors to publish data, computer code, and computational environments alongside their publications to increase reproducibility. This puts yet another pressure on faculty, which is apparent in multiple surveys that ask researchers about their experiences. Increased workload and time requirements, a lack of formal training in ensuring reproducibility, and the lack of support tools are often identified as barriers (see e.g., Baker, 2016 or Konkol et al., 2019). Institutional data repositories such as the FIU Research Data Portal described in this paper can help offset this extra workload and time. Compared to third-party repositories similar in functionality (e.g., Zenodo, Figshare), an institutional repository with its supporting infrastructure (dedicated staff, tutorials, etc.) is more efficient in providing this support. At FIU, workshops, FAQs, and user guides are available to faculty.

Discussion and Conclusion

Library-based research data management and curation often faces multiple challenges, including lack of human resources in both FTE commitment from the university and expertise needed for this complex matter. Other challenges include strategic vision for library-based RDM, and fiscal limitations, which often lead to sustainability issues, as long-term preservation of institution-generated datasets could be very costly. Innovation and collaboration are essential to the successful implementation of a library-based RDM. Furthermore, a useful library-based RDM must go beyond the concept of metadata management for datasets. It needs to be connected with other academic data services, technological tools, training, and consultation services. Working with domain experts to understand their data management needs and user cases is an essential part of the RDM. Moreover, data specialists and data librarians should be equipped with data science knowledge, which is different from information science. In the era of big data, libraries are also tapping into text-mining and the AI arena. Libraries' digital collections services, digital humanities, digital scholarship, digital initiatives, institutional repository systems, and geographic information systems (GIS) can all be connected to digital and data science, for which research data management (RDM) can serve as a foundation. Future academic libraries will inevitably expand into the universe of data and its related services.

References:

Akers KG, Doty J (2013) Disciplinary Differences in Faculty Research Data Management Practices and Perspectives. *The International Journal of Digital Curation,* 8(2). http://www.ijdc.net/article/view/8.2.5

CARDH, https://cardh.org/.

CRIES, https://www.cries.org/.

Team, DataCite. *DataCite*, https://datacite.org/.

The Dataverse Project - Dataverse.org, https://dataverse.org/about.

DataONE, 13 June 2022, https://www.dataone.org/.

De Rond, Mark, and Alan N. Miller. "Publish or Perish." *Journal of Management Inquiry*, vol. 14, no. 4, 2005, pp. 321–329., https://doi.org/10.1177/1056492605276850.

Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., Olds, J. L., & Dean, H. (2015). Social, behavioral, and economic sciences perspectives on robust and reliable science. *Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences*, *1*.

Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533,** 452–454 (2016). https://doi.org/10.1038/533452a

Markus Konkol, Christian Kray & Max Pfeiffer (2019) Computational reproducibility in geoscientific papers: Insights from a series of studies with geoscientists and a reproduction study, *International Journal of Geographical Information Science*, 33:2, 408-429, DOI: 10.1080/13658816.2018.1508687

"Gallery." *ArcGIS*, ESRI, https://doc.arcgis.com/en/hub/gallery/.

"Living Atlas of the World." *ArcGIS*, https://livingatlas.arcgis.com/en/home/.

Florida International University Geoportal, http://dpanther2.fiu.edu:8080/geoportal/catalog/main/home.page.

Florida International University Libraries Research Data Portal, http://rdm.fiu.edu/.

Florida International University Security Research Hub, https://srh.fiu.edu

Hudson-Vitale, Cynthia; Imker. "SPEC Kit 354: Data Curation." *Association of Research Libraries Digital Publications*, 9 May 2017, https://publications.arl.org/Data-Curation-SPEC-Kit-354/.

Knb.ecoinformatics.org, https://knb.ecoinformatics.org/data.

Make Data Count, 26 May 2022, https://makedatacount.org/.

"Metadata in the ArcGIS Ecosystem." *Metadata-ArcGIS Online Help | Documentation*, https://doc.arcgis.com/en/arcgis-online/manage-data/metadata.htm#ESRI_SECTION2_9AB0CCA6A1C443C5A0AEA956D15C1E55.

"Metadata References." *Dataverse.org*, https://guides.dataverse.org/en/latest/user/appendix.html#user-appendix.

Piwowar, Heather A. "Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data." *PLOS ONE*, Public Library of Science, https://doi.org/10.1371/journal.pone.0018657.

Yoon, Ayoung, and Devan Ray Donaldson. “Library Capacity for Data Curation Services: A US National Survey.” *Library Hi Tech*, vol. 37, no. 4, 2019, pp. 811–828., https://doi.org/10.1108/lht-12-2018-0209.