Journal of eScience Librarianship

Volume 8
Issue 2 *Special Issue: Research Data and Preservation (RDAP) Summit 2019*

Article 8

2019-12-23

Data Curation for Big Interdisciplinary Science: The Pulley Ridge Experience

Timothy B. Norris
University of Miami

Et al.

ISSN 2161-3974

JeSLIB 2019; 8(2): e1172
doi:10.7191/jeslib.2019.1172

Journal of eScience Librarianship
putting the pieces together: theory and practice

Full-Length Paper

Data Curation for Big Interdisciplinary Science: The Pulley Ridge Experience

Timothy B. Norris and Christopher C. Mader

University of Miami, Coral Gables, FL, USA

Abstract

The curation and preservation of scientific data has long been recognized as an essential activity for the reproducibility of science and the advancement of knowledge. While investment into data curation for specific disciplines and at individual research institutions has advanced the ability to preserve research data products, data curation for big interdisciplinary science remains relatively unexplored terrain. To fill this lacunae, this article presents a case study of the data curation for the National Centers for Coastal Ocean Science (NCCOS) funded project "Understanding Coral Ecosystem Connectivity in the Gulf of Mexico-Pulley Ridge to the Florida Keys" undertaken from 2011 to 2018 by more than 30 researchers at several research institutions. The data curation process is described and a discussion of strengths, weaknesses and lessons learned is presented. Major conclusions from this case study include: the reimplementation of data repository infrastructure builds valuable institutional data curation knowledge but may not meet data curation standards and best practices; data from big interdisciplinary science can be considered as a special collection with the implication that metadata takes the form of a finding aid or catalog of datasets within the larger project context; and there are opportunities for data curators and librarians to synthesize and integrate results across disciplines and to create exhibits as stories that emerge from interdisciplinary big science.

Correspondence: Timothy B. Norris: tnorris@miami.edu
Keywords: Data Curation, Big Science, Data Repository, Story Maps, Interdisciplinary

Disclosures: The substance of this article is based upon a poster presented at RDAP Summit 2019. Additional information at end of article.

Introduction

The curation and preservation of scientific data has long been recognized as an essential activity for the reproducibility of science and the advancement of knowledge. For nearly a half century investment into, and development of, data curation for big science in specific disciplines and at individual research institutions has advanced the ability to preserve research data products. The Long Term Ecological Research Network (LTER), the Inter-Institutional Consortium for Political and Social Research (ICPSR), NASA's Global Imagery Browse Services (GIBS), and the National Human Genome Research Institute (NHGRI) serve as emblematic successful outcomes of these efforts. Data curation for interdisciplinary big science remains relatively unexplored terrain, however.

Interdisciplinary Big Science, for the purposes of this paper, is understood as well-funded science undertaken by inter-disciplinary teams of researchers with the intent to better understand wicked problems that are not located within a single disciplinary realm and cannot be solved neatly from one perspective, if at all (Rittel and Webber 1973). Such interdisciplinary big science approaches to wicked problems are gaining increased attention and funding at institutional and national levels. As examples: the University of Miami Laboratory for INtegrative Knowledge (U-LINK) and the University of Texas at Austin Bridging Barriers program both support such research at institutional levels; the National Science Foundation Smart and Connected Cities is a cross-directorate initiative that encourages interdisciplinary and inter-institutional work. Many other contemporary examples can be identified.

As these interdisciplinary problem driven programs gain traction, researchers continue to publish results and curate data within disciplinary silos, but struggle to find tools, mechanisms, and incentives to publish research results that address the wicked problem identified at the outset of the project. This challenge for researchers translates to an opportunity for data curators and librarians. With this opportunity identified as such, this paper reports on an embedded data curation experience within interdisciplinary big science.

In 2011 the University of Miami Center for Computational Science (CCS) was enlisted to collaborate as data curators on a multi-year interdisciplinary NOAA National Center for Coastal Ocean Science (NCCOS) funded research project located in the Gulf of Mexico. Specifically the CCS was enlisted to build a Decision Support Resource (DSR). The DSR was envisioned as a multi-tier, web-based software application that provides comprehensive access to data and analyses generated by the project. It is comprised of three basic components: a metadata store, a repository, and data exploration tool. The construction of the DSR was an intentional exercise in data curation for interdisciplinary big science.

After a brief literature review for data curation with a specific focus on data sharing and publication, this article reports on the technical, social and political aspects of Pulley Ridge data curation experience. An introduction to the NOAA/NCCOS funded project gives context to a description of the methods used to build the decision support resource. The methods are described with a focus on the metadata store which leads to an outline of the repository architecture and functionality. A discussion of project successes, failures and lessons learned in terms of technical, social and political aspects follows. Drawing from this reflection a case is made that data from interdisciplinary big science can be thought of as a special collection. The article closes with remarks on the implications of considering data as a special collection and

presents potential ways forward for data curation for interdisciplinary big science.

Data Curation and Data Sharing in the Academy

Data curation is typically defined as a set of activities that generally add value to data. For example, Giaretta defines digital curation as "maintaining and adding value to, a trusted body of digital information for current and future use" (2008). Curation activities can include selecting and maintaining bodies of data, annotation, building linkages or interoperability, management of large data sets, data validation, editorial input, archiving, preservation, and so on (Beagrie 2008). The Data Curation Network recently identified 47 data curation activities with the top five researcher identified activities as: data documentation, preserving the "chain of custody," secure data storage, quality assurance for data, and minting of persistent identifiers (Johnston et al. 2018). From a practical standpoint, the definition of data curation as a set of activities serves both data curators and data creators in the shared process to identify and refine best practices. The list of activities also serves to emphasize that data curation within an academic context often leads to a some form of data publication or data sharing.

Long-term research on scholarly communication and data sharing by scientists shows several common barriers that make data sharing among scientists difficult. The barriers include: time and resources necessary to prepare well organized data packages and to create quality metadata to describe the packages; loss of control of how data is used once it has been shared; minimal recognition among peers for sharing data – especially in the tenure review process; the career trajectory of the researcher; complex ownership and licensing issues; and the discipline of the research (Tenopir et al. 2011; Fecher, Friesike, and Hebing 2015; Tenopir et al. 2015; Berghmans et al. 2017; Stuart et al. 2018). This list is by no means exhaustive but clearly shows a connection between data curation activities and data sharing by scientists. Research also shows that there are positive trends in data sharing practices, particularly over the last decade (Tenopir et al. 2015).

It is important that these approaches to, and analysis of, data curation in an academic context draw from historical experience with the curation of museum exhibits and other types of collection curation (for example see Fry 1965). In these broader contexts there is often an overarching goal to educate or to tell a story, or in the words of Fry, the "Presentation and Dissemination of the Results" (Fry 1965 p. 245). On the other hand, with data curation in the academy, overarching goals focus on findability, accessibility, interoperability and reproducibility; the concepts outlined as FAIR data (Wilkinson et al. 2016). The goal to disseminate or to tell a story is left to separate forms of communication, scholarly or otherwise, and is not explicitly included. This article speaks to this difference in purposes and an argument is made to give a higher priority to the presentation and dissemination of research results in data curation work.

Project Background: Big Interdisciplinary Science at Pulley Ridge

In 2011 the CCS was enlisted to collaborate as data curators on a multi-year trans-disciplinary project titled "Understanding Coral Ecosystem Connectivity in the Gulf of Mexico-Pulley Ridge to the Florida Keys." The project itself studied how the benthic formation at Pulley Ridge connects to the ecosystems of the Florida Keys National Marine Sanctuary and other socio-natural communities in the Caribbean and South Florida regions. For the purpose of the

JeSLIB 2019; 8(2): e1172
doi:10.7191/jeslib.2019.1172

grant narrative, the management of the natural resources in the project area comprises a wicked problem; the problem definition implies the problem solution, a solution for one stakeholder may be a problem for another stakeholder, there is no problem resolution but instead an improvement of the situation, and so on (c.f. Rittel and Webber 1973). The knowledge generated from the project is intended to inform decision making for management, conservation and protection of the Pulley Ridge benthic formation and its associated biological community. This management problem includes stakeholders within the south Florida industries of fishing and tourism and the sustainable management of the Florida Keys.

The project "represented a collaboration of more than 25 scientists at nine different universities and two federal laboratories pooling their expertise through NOAA's Cooperative Institute for Marine and Atmospheric Studies at the University of Miami in coordination with the Cooperative Institute for Ocean Exploration, Research, and Technology at Florida Atlantic University" (NCCOS n.d.a). At the outset the project was ambitious; the grant narrative crossed disciplinary boundaries spanning the physical, biological and social sciences. With this planned complexity, the NOAA/NCCOS program officer overseeing the grant review process suggested to the three principal investigators on the project that there be a data curation component. This program officer initiative led to CCS enlistment and involvement.

At the outset of the project one concrete deliverable from the CCS was to identify and produce a set of Decision Support Tools (DST). CCS involvement during the first several years of the project was primarily limited to participation in an annual meeting, as well as other occasional meetings organized to communicate design concepts for a possible DST. During these first several years it was agreed upon that the ultimate form of the DST could better be described as a Decision Support Resource (DSR), that would help interested parties (including natural resource managers and stakeholders) understand the types of data created through the project by the individual investigators (as opposed to supporting specific decision making activities). Ideally, the DSR would also help these interested parties to understand some big picture integrative analyses that might say something about the connectivity of the several ecosystems and their social counterparts studied during the project. At this point the design focus turned toward data curation and access, as well as support for some kind of integrative story telling.

The initial data curation team was comprised of four dedicated data curators from the University of Miami CCS, one metadata/cataloging specialist at the NOAA/NCCOS with close ties to the National Center for Environmental Information (NCEI), and the program manager at NOAA/NCCOS. The CCS team was comprised of three software engineers and a geographic information systems (GIS) intern. In 2015 an additional curator joined the Miami team; a Council of Library and Information Resources (CLIR) postdoctoral fellow with geospatial data experience. All of the data curation was collaborative and, in practice, experimental.

The Decision Support Resource

The DSR was developed to provide comprehensive access to the scientific data and analyses generated by the project. In some ways it duplicates functionality available from other tools. For example, open source tools such as GeoBlacklight (https://geoblacklight.org), the Knight Labs storymap.js project (https://storymap.knightlab.com), and Omeka (https://omeka.org) with the Neatline plugin (https://neatline.org) strive to provide geospatial repository functionality and

JeSLIB 2019; 8(2): e1172
doi:10.7191/jeslib.2019.1172

online geospatial exhibits respectively. Currently the open source Spotlight project at Stanford (https://library.stanford.edu/research/spotlight) likely provides the most robust combination of repository and storytelling. Additionally, the proprietary ArcGIS online and ArcGIS story map products are designed for geospatial data sharing and online geospatial data exhibits. Nevertheless, there is no open source solution that meets the dual purpose of a geospatial data repository and a geospatial story telling platform in a satisfactory way.

The DSR includes a novel technological approach to building such systems and leverages CCS team experience on other science consortium data management projects (Mader et al. 2015, Center for Computational Science 2015). The DSR is a publicly available web-based application built on the MEAN server stack (MongoDB, express, angularJS, and Node.js), written in Java, Javascript, HTML and CSS, and makes use of Apache Solr to provide flexible and rapid search capabilities. It is comprised of three basic components: a metadata store, a repository and data exploration tool. The current iteration of the DSR enables users to interact with project data and analyses in three specific ways: 1) exploring data sets; 2) exploring data set layers; and 3) viewing interactive data exhibits (see Figure 1). The DSR is now available to the research team, resource managers, legislators, and the private sector (Center for Computational Science 2018).

DSR Component		Facilitates
metadata store	→	data set exploration
repository	→	data exploration
data exploration tool	→	interactive visualization

Figure 1: Components of the Decision Support Resource

From the outset the DSR was planned to compliment data deposits in discipline specific repositories such as the National Center for Environmental Information (NCEI) and The National Center for Biological Information (NCBI). Early in the project a working relationship was established between the CCS and the NCCOS/NCEI to ensure metadata interoperability. For data sets being deposited at NCEI, the CCS team helped to assemble the data packages for deposit. Data sets based on genomic sequence information were generally deposited directly by the researchers to the NCBI without assistance.

Metadata Store

Initial work focused on creating several layers of metadata; one set for the entire project and subsequent descriptions for each dataset within the project. The DSR was designed to work with a specific data model and conceptual process model for registration of the data sets accessible through the tool. Core to the concept was the identification of metadata about the datasets, as well as extraction of data summaries for exposition (i.e., map-based display). To gather this metadata each individual researcher was asked to complete a document-based form to help capture a good description of their specific disciplinary work (word document). The

JeSLIB 2019; 8(2): e1172
doi:10.7191/jeslib.2019.1172

metadata form was drafted as a collaboration between the CCS and the NCCOS/NCEI. Initially the CCS chose one set of Dublin Core-based metadata elements focused on project level description, and a secondary set of elements drawn from the ISO 19115 metadata standard focused on descriptions specific to discipline-based datasets to be captured throughout the research effort. This first draft was then integrated with the existing metadata elements used by NOAA/NCCOS for data archiving. The final draft of the metadata template is available as a supplemental file to this article.

Perhaps unsurprisingly, providing metadata descriptions for data during the collection process was not a high priority for researchers. To resolve this difficulty, what became known as "researcher dunning" was carried out by email for a period of nearly two years. The dunning process consisted of persistent email requests for metadata from the researchers. Often responses were not forthcoming and the email requests were simply repeated until a response was received. This approach resulted in an iterative process between the CCS curation team, the NOAA/NCCOS program manager, the NCEI, and the researchers themselves to select and refine metadata elements (see Figure 2). The result is a refined set of metadata elements that can be used for every research product within the entire project. In total, 28 rich metadata records were produced for all of the shareable data produced from the research.

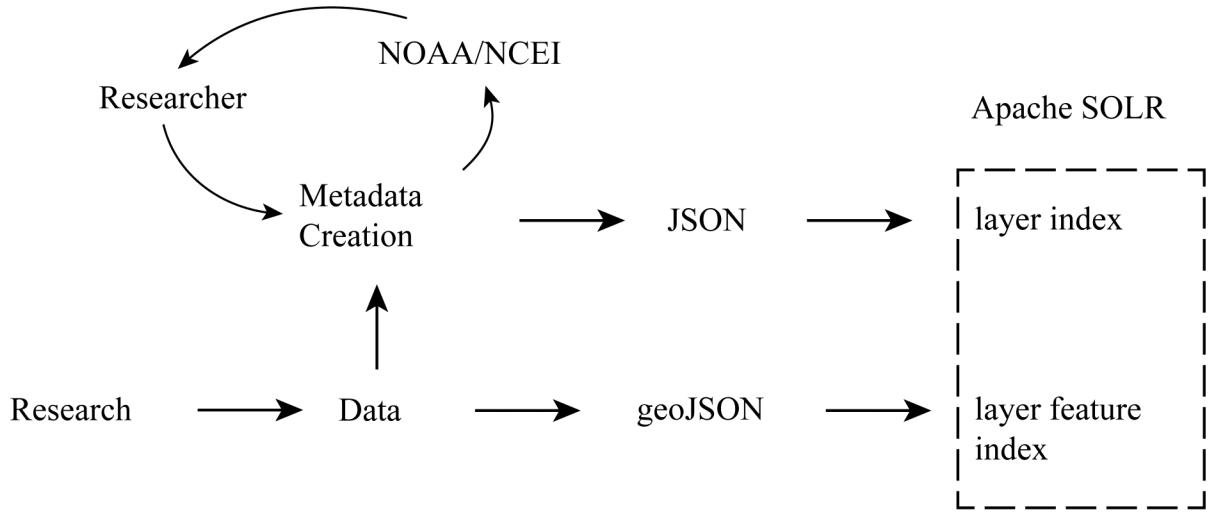

Figure 2: Workflows for the creation of the metadata and the two SOLR indices used in the DSR application.

Exploring Data Sets and Layers – the Repository

Parallel to the writing of the metadata, an indexing system built on Apache SOLR technology was created that ingests metadata formatted as plain text in JavaScript object notation (json). This index enables the implementation of search functionality for the front-end DSR web application. A second index was built that ingests actual data points from the research in the field, lab and modeling systems. This second index requires geospatial coordinates which are then used for geographically visualizing the dataset in question. It only ingests data represented as GEOjson; in those cases where no geospatial data existed in the dataset,

JeSLIB 2019; 8(2): e1172
doi:10.7191/jeslib.2019.1172

bounding boxes (for geospatial raster datasets) and links to non-geographic data were created. Thus, the front-end web application can access the two indexes through a java api which returns a reference to the dataset, and when available, the georeferenced dataset itself (see Figure 2).

The repository component of the DSR built by the CCS was envisioned as a compliment to discipline-specific repositories. As examples, the ships log data from the annual cruises, the data from the benthic surveys and data from the genetic component of the project are housed at the NCCOS, the NCEI or the NCBI, respectively (NCCOS n.d.b). Within the DSR, download links to these repositories are provided for published, publicly available, data sets: currently the NCEI and the NCBI, but other sites can be accommodated as needed. In this sense the DSR is an umbrella repository that brings together metadata for the entire project, houses some of the data directly, and provides access to data housed elsewhere.

The repository interface provides a free text search, faceted filtering, data download links and citation tools. DSR users can search for data sets and filter results using a combination of free text search and faceted filtering. These search paradigms draw from ecommerce web site design. For example, a user may search for all data sets referencing “Montastraea” and then filter those results to only those data sets related to “Population Genetics.” Or as another example, a user may want to filter all datasets for those collected in situ and for those that that have been submitted to the repository (see Figure 3). Each data set in the DSR is also provided with a citation that can be simply copied and pasted, or downloaded in a number standard metadata formats (e.g., EndNote or Bibtex). DOIs (where available) are provided for each data set.

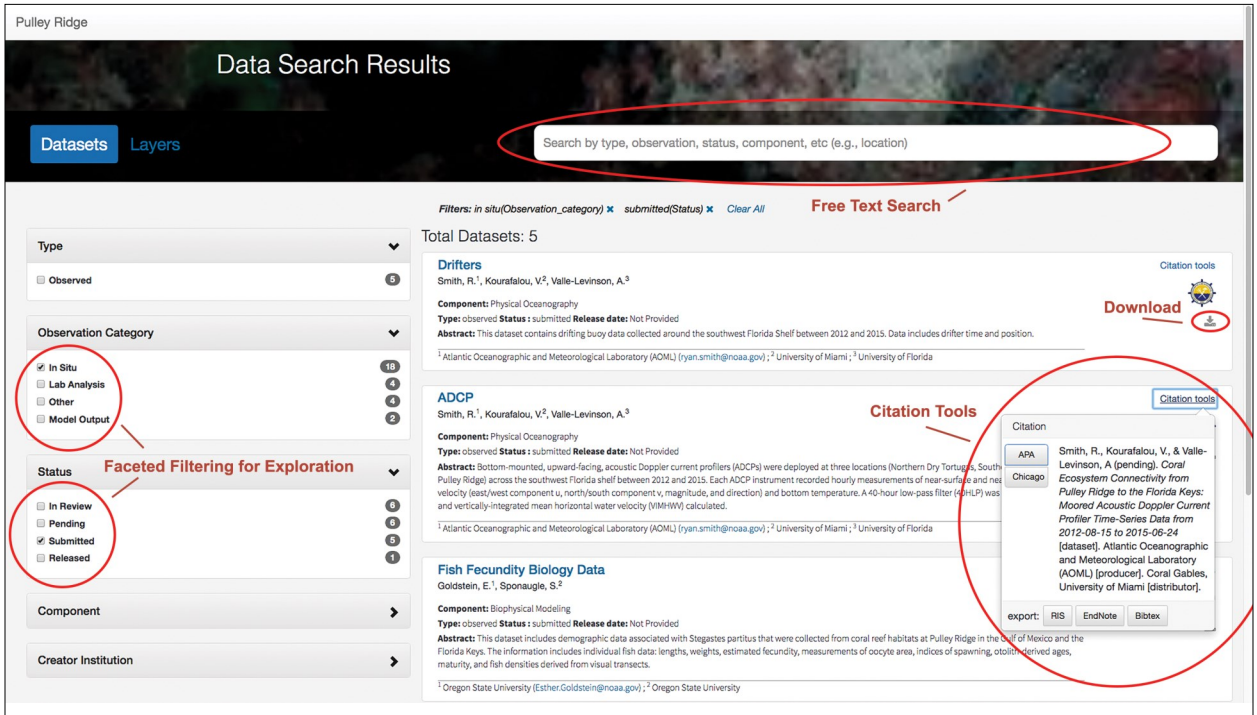

Figure 3: Screenshot of the repository showing the ability to perform full text searches, faceted filtering, download citation information and download the data itself.
https://mesophotic.ccs.miami.edu/discover/datasets

JeSLIB 2019; 8(2): e1172
doi:10.7191/jeslib.2019.1172

And finally, each data set is presented, when relevant, with a set of "layers" displayed on a map-based interface (see Figure 4). These layers are georeferenced and show the geographic location of the data set both as the data set extent and as the specific locations of the observations or model output. For example, the Drifter data set contains "drifting buoy data collected around the southwest Florida Shelf between 2012 and 2015. Data includes drifter time and position" (Smith, Kourafalou, and Valle-Levinson *pending*). The DSR enables users to preview the extent, which covers several thousand square miles, and also to see the approximate locations where the drifters were released and their paths across the ocean surface (see Figure 4).

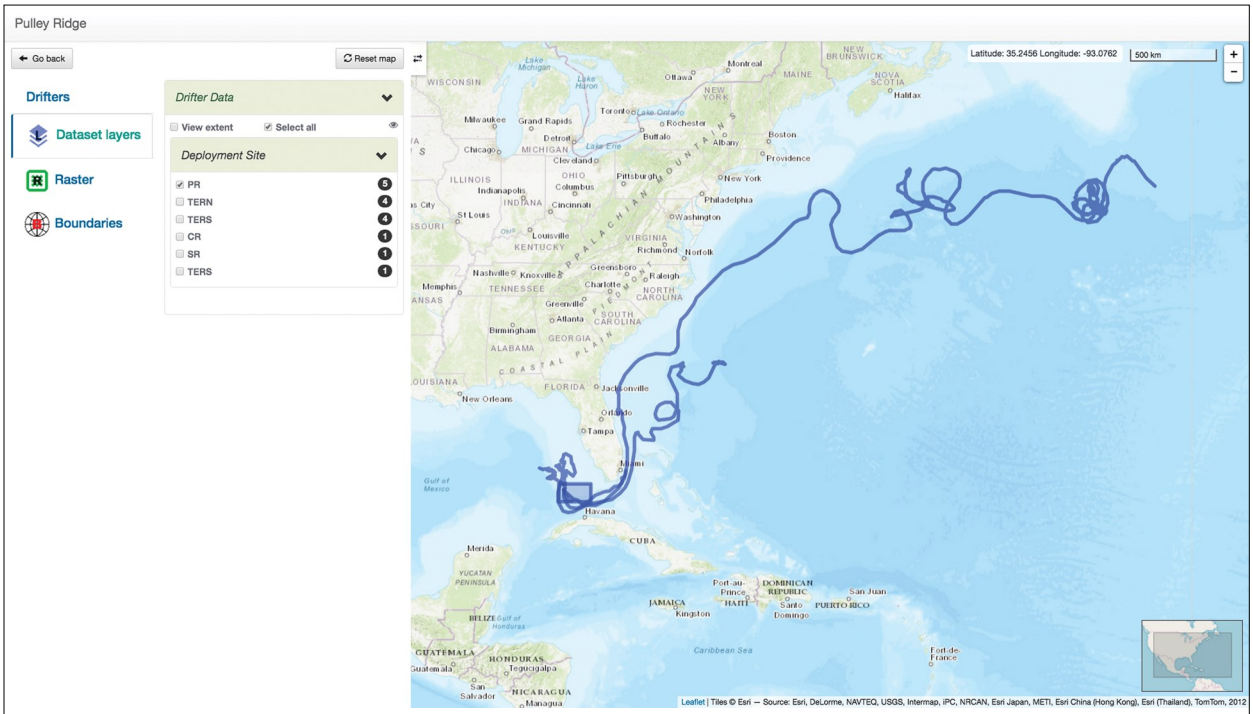

Figure 4: Screenshot of the repository showing the ability to quickly visualize the geospatial extent and characteristics of individual datasets.
http://mesophotic.ccs.miami.edu/mapViewer/layers_collection/5a9405fb1fe9dd27205fc566/layers

Interactive Maps and Geospatial Data Visualization

The storytelling tool provides deeper context for data from the project. Four "scenarios" were planned as a way to categorize and synthesize project output: Genetic Connectivity; Bioeconomics; Physical Dynamics; and Biodiversity. Each scenario presents project data in a narrative format as a series of "slides" containing maps, charts and narrative explanation—a story map. The goal of the scenario sections is to tell a thematic story about Pulley Ridge, the Gulf of Mexico, and the Florida Keys by synthesizing results and data produced by allied disciplines from within the project. For example, the "Bioeconomics" scenario presents fisheries economic value and fishing catch logs in the context of the Eastern Gulf of Mexico and Florida Keys, highlighting the time period prior to the establishment of the protected area the Pulley Ridge Habitat Area of Particular Concern (HAPC), through the present, as well as presenting several future options for policy changes to the HAPC (Die et al. 2018).

JeSLIB 2019; 8(2): e1172
doi:10.7191/jeslib.2019.1172

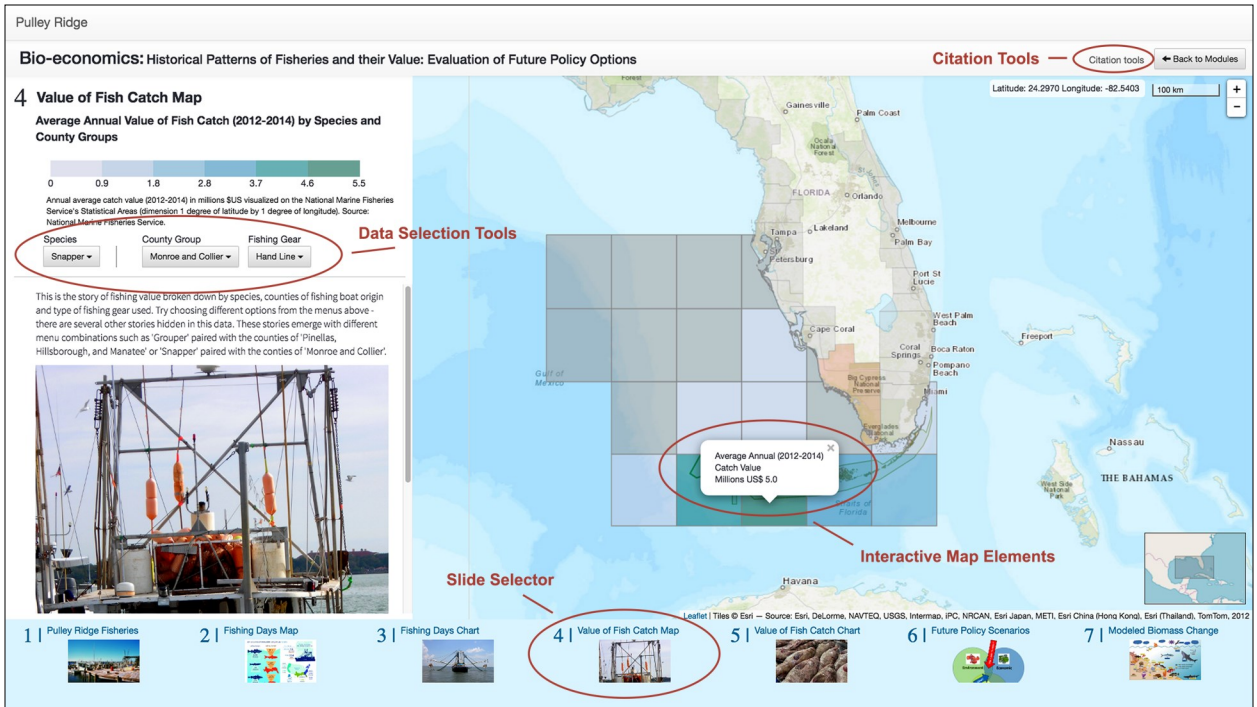

Figure 5: Screenshot of the storytelling tool showing a slide from the bioeconomic story: a map of summarized fishing catch data. See Figure 6 for an alternate visualization.
https://mesophotic.ccs.miami.edu/explore/module/bio-economics/3

An example “slide” from the Bioeconomics scenario presents the “Average Annual Value of Fish Catch (2012-2014) by Species and County Groups” as an interactive map and narrative (see Figure 5). Users may explore the value and location of the commercial catch for any combination of several species, gear types and county-based fleets. Through the interactive tools on this slide, the story emerges that the counties of Monroe and Collier almost exclusively fish in Pulley Ridge and Keys, whereas Pinellas, Hillsborough and Manatee counties mainly focus their fishing effort north of those areas. The scenarios also have a citation tool and can reference published articles and users can navigate back to supporting data sets and their metadata in the repository (see figures 5 and 6).

As originally planned, this component of the DSR was to be administered by the researchers themselves. The stories would emerge through the DSR facilitated ability of the researchers to visualize and overlay multiple datasets from individual studies carried out in different disciplinary contexts. Slides would be created by drawing on multiple datasets, synthesizing

and integrating the results, adding a narrative, and then linking several slides together in a story. Within the DSR data sets can be selected for use in a slide, geographic visualization styles can be set, and interactive functionality can be added. On the server back end, the slides that made up a story map are stored as a json document in a MongoDB database. In the front end web application the story map was rendered with the leaflet.js and chart.js JavaScript libraries for mapping and graphics respectively. While the technical aspects of the story map functioned more or less as expected, the researcher input for the creation of the integrated stories only occurred for the bio-economic scenario.

JeSLIB 2019; 8(2): e1172
doi:10.7191/jeslib.2019.1172

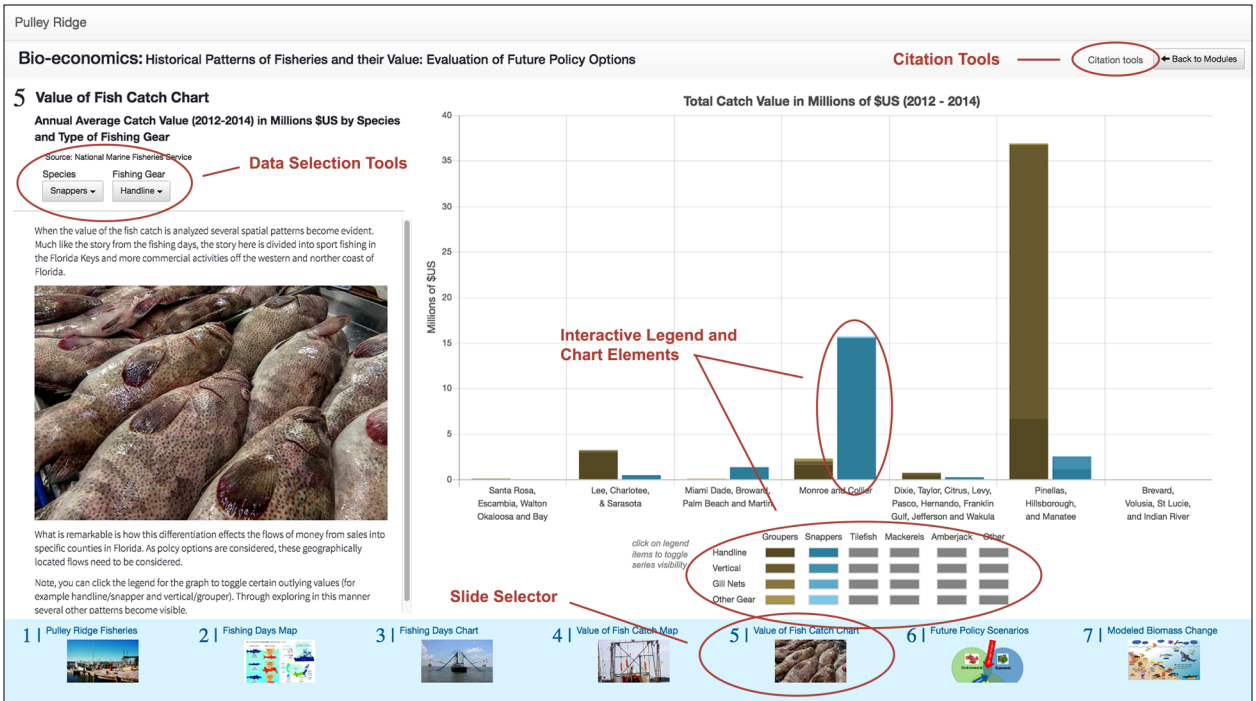

Figure 6: Screenshot of the data exploration tool displaying an interactive chart to visualize the same fishing catch data in Figure 5. https://mesophotic.ccs.miami.edu/explore/module/bio-economics/4

Discussion

The creation of the DSR as a data curation experience for big interdisciplinary science was experimental. The combined interest from NOAA/NCCOS, the principal investigators, and the CCS in conducting this work allowed for creativity, exploration and cross-fertilization between these groups. The construction from the ground up of the repository and data exploration tool provided a forum in which software developers, database mangers, researchers, and data curation experts could all come together to present their distinct and not-always-aligned approaches to the table. Furthermore, this confluence of interest coincided with a broader trend in US government open data which crystalized in the 2013 Office of Science and Technology Memo for “Increasing the Access of the Results of Federally Funded Scientific Research” (Holdren 2013). The timing of the OSTP memo helped to energize the work. The discussion that follows draws from this experience.

Re-implementation of the Repository Wheel

The DSR is built using a set of modern, widely used, open source components and libraries that undergo regular revision and modification, as well as code developed at the UM CCS. Open source components include the MEAN server technology stack, leaflet.js, node.js, PostgreSQL, and Apache Solr. Our decision to use open source software was based primarily on the belief that systems developed using these technologies have the potential for greater sustainability and reuse over time than those developed using closed source technologies (e.g., ESRI). In the case at the CCS, the DSR represents the vanguard of a set of web-based geospatial systems, including other ongoing work with Miami-Dade County

JeSLIB 2019; 8(2): e1172
doi:10.7191/jeslib.2019.1172

(https://land.ccs.miami.edu), with informal communities in Colombia (participatory mapping with the community of Las Flores), and the City of Miami (http://healthtool.ccs.miami.edu/HealthFoundation/map). The DSR repository and data exploration tool was intentionally built as an online GIS that integrates with this other work. While the specific future of the DSR itself is unclear, the technology and experience gained during the development process currently lives on as a shared resource between the CCS and the University of Miami Libraries. While this resource is of incalculable institutional value, several shortcomings of the DSR are now evident.

As the repository was built, no one person on the team had deep knowledge of OAI-PMH protocols and standards (https://www.openarchives.org), and with such a knowledge gap, the DSR does not provide an API for metadata harvesting using the OAI-PMH protocol (or any other protocol). From a library and institutional perspective, the discoverability that OAI-PMH facilitates is invaluable and the lack of this functionality makes the DSR a less desirable repository solution. This technology can be relatively easily implemented in a future version, nevertheless.

The repository component was designed and built around a geospatial data model that enables the Apache Solr engine to best index data for discovery, flexibility with kinds of data that can be deposited, and appending geospatial attributes to non-geospatial forms of data. The interactive storytelling component of the DSR incorporates the same data model into an extended metadata model to represent maps as collections of distinct datasets within the repository (data from the project) and data from outside of the repository (reference maps and base maps). This integration proved difficult as each map is a unique combination of data layers from the project, base data layers and geometries, and interactive functionality. As examples: time base animations where the time dimension is in the project data layer and the geometry is a base map layer; or the categorization of project data based on field values where unique categorization schemas are needed for each data layer. These are classic cartographic problems. The technological aspects are relatively easy to address in future versions of the DSR, but the human based cartographic process will always be challenging (more below).

Metadata Creation and Data Description

The iterative process to create a metadata standard for the entire project highlights the role of the metadata as a finding aid. The data that emerged from the Pulley Ridge project can be seen as a special collection as understood by library and information science professionals, and the DSR as a combination of the metadata repository and the visualization tools is an online exhibit. The data itself is a unique heterogenous acquisition for the research institution that requires significant work to catalog and ingest. Once ingested a finding aid must be created to make it useful to researchers that may or may not be familiar with some of the disciplinary norms used for the data collection and data documentation. And finally, an exhibit to showcase the special collection increases the collection's overall value and impact. The significance of this should not be overlooked as there are several opportunities within this observation.

First, and in no way to diminish the "Special Collections as Data" conversation (Padilla et al. 2019), we can reverse the words and speak of "Data as Special Collections." While this does

not apply to the long tail of scientific data as outlined by Heidom (2008), approaching data curation for interdisciplinary big science from this perspective may help identify gaps in such curation processes. For example, we observed that on the one hand the curation team did have the required dedication to the subject matter for cataloging the special collection. On the other hand the curation team has little to no experience with library special collections and may have overlooked important aspects of creating a finding aid. In any case, there is likely much to learn from library special collections that can be applied to curating data as special collections.

Second, from a data as special collection perspective, the goal of data curation expands to become more than simply meeting the FAIR principals for data publication (Wilkinson et al. 2016), but instead takes on a form of novel scholarly communication. The purpose of curating the data expands to include the political process of sharing the information with communities who might benefit from the accessibility of the data and the underlying story to be told. In the current political climate of a “post-truth society” where scientific results rarely reach the general public, and when they do are often distorted (Iyengar and Massey 2018), novel forms of scientific and scholarly communication can help remedy this situation. As part of the process to curate data as a special collection, humans intentionally use data to tell stories as opposed to simply letting data tell stories with no human intervention. Often the creation of an exhibit from a library special collection is an intentional political act, perhaps the same may be true for an exhibit created from a special collection of data.

Researchers, Data Sharing and Scholarly Communication

Almost all data sharing obstacles mentioned in the literature were observed throughout the curation process. Similar to the difficulties encountered in capturing good metadata from the researchers, regular researcher dunning was also undertaken to get the actual data products from the research team. Even after persistent repeated email requests for research data products, to date only five datasets have complete submissions. For comparison, 25 peer-reviewed publications have emerged from the research. We interpret this data sharing difficulty as a researcher response to the lack of incentive to share data once the journal article is published. Dedicating time to the next upcoming research project always wins over giving time to publish an already used dataset that returns little to no recognition in the tenure process. The exception was the genomic sequence data which was directly deposited by the researchers to the NCBI without assistance.

Within the funding period of the project, researchers did not work together to synthesize results and tell integrated stories to present findings that crossed disciplinary boundaries. The story of the interconnectivity of Pulley Ridge, with physical, ecological, and economic systems across its peripheral regions is what drove the grant narrative. This included application of the results to the political and social processes—the implementation of conservation measures and mitigation plans for the Pulley Ridge formation—involved in managing the natural resources of the region (the wicked problem). Yet when it came to move the project results into the political and social realm through non-scholarly communication, the research team was unable to dedicate sufficient time to this process. This can be attributed to the difficulty of the (cartographic) task, the time needed to create such integrative stories, and that the researchers understandably were focused on their own discipline bound research. Perhaps most important, the storytelling idea came from the NOAA/NCCOS program officer and not the grant proposal itself.

The bioeconomic scenario is an exception, but even for this integrated story the work to synthesize results from distinct research components was mostly undertaken by the CCS curation team with limited guidance from the three researchers involved. To organize and lead this effort was an unplanned and unexpected role for the CCS team. In this case, once the results were synthesized and incorporated into an interactive story map, the researchers were intrigued by the novel presentations of distinct datasets as one story. While the work necessary to produce the story map provided an important space for three researchers to come together with the curation team and discuss their results, it was the interactive presentation that gained their attention. The researchers had never seen the distinct datasets presented side by side in a dynamic, integrated and interactive presentation. The process to integrate results made visible stories that neither the researchers nor the data curators had considered previously.

Conclusion: Dissemination of Integrated Results in Interdisciplinary Big Science

Several conclusions relevant to e-science librarianship emerge from the Pulley Ridge Data Curation Experience. Data from interdisciplinary big science is similar to a library special collection; this is an opportunity to create metadata as finding aids and to build purposive exhibits to add value and increase impact. Furthermore, curating data as a special collection can dovetail with the current efforts to create institutional data catalogs for research universities. For this purpose, and in some cases, geographic information systems (GIS) may be the ultimate data curation tool; a well-constructed GIS is a *special collection of data* made purposefully to *curate cartographic exhibits*. Nevertheless, from a technical perspective, most GIS tools lack certain library information science functionality such as implementations of the OAI-PMH protocol for harvesting metadata. Additionally, the creation of maps as exhibits is time consuming and rife with cartographic difficulties. Finally (and significantly), construction of the DSR for the Pulley Ridge project built institutional knowledge and community around data curation at the University of Miami. This knowledge is already in use for ongoing data curation work.

Perhaps the most significant conclusion is that the lack of planning for post-project data curation in interdisciplinary big science is an opportunity for data curators to become data synthesizers, integrators of project results, and ultimately storytellers. Often research proposals that outline interdisciplinary approaches to wicked problems, little planning exists for the work necessary to integrate the final results from distinct disciplines, to use the integrated results to communicated findings either within the project or to decision makers and the general public, and ultimately to address the wicked problem as described in the original proposal. This lack of planning spans institutional to national levels, such as the U-LINK program described in the introduction and the Pulley Ridge Project described in this article, respectively. As a result, there are no resources set aside to perform this work, there is no overarching data management plan that allows for data interoperability within the project, and there are few incentives for disciplinary researchers to take on the task of integrating and communicating results to others outside of their discipline.

This observation highlights a point of intervention in project planning and grant writing for data curators, particularly for those who have experience with the curation of special collections, with the curation of exhibits, and those with data curation experience. Perhaps there never will be incentives for discipline-based researchers to integrate and synthesize data across

disciplinary boundaries in interdisciplinary projects, but through careful planning resources can be made available to librarians and data curators as academic generalists to perform this work. While this may not be a standard or planned role for data curators or librarians, the recognition of this opportunity may help resolve the social difficulty of academic data sharing and the political difficulty of non-scholarly communication for applied research. To move in this direction more research is required to better understand how data from big interdisciplinary science can be curated as a special collection. Additionally new hires in academic data curation can be aligned with interdisciplinary big science to facilitate data sharing and non-scholarly communication that stems from this kind of applied research and thus better address the wicked problems that we face as a society.

Acknowledgements

Many thanks to Dr. Robert Cowen (OSU, previously at UM), Dr. Peter Ortner (UM – RSMAS), and Dr. Shirley Pomponi (FAU – HBOI), the three principal investigators on the Pulley Ridge Project. This work was funded by the NOAA National Centers for Coastal Ocean Science awards NA11NOS4780045, NA09OAR4320073, and NA14OAR4320260. We would also like to thank Kimberly Puglise and Jessica Morgan of the NOAA National Centers for Coastal Ocean Science. Special thanks to Julio Perez (software developer), Sreeharsha Venkatapurum (databases and indexing) and Chance Scott (Geographic Information Systems) of the University of Miami Center for Computational Science. Without their expertise this project would not be possible.

Disclosures

The substance of this article is based upon a poster presented at RDAP Summit 2019: “The Pulley Ridge Data Curation Experience” available at https://osf.io/6t4c5.

References

Beagrie, Neil. 2008. "Digital Curation for Science, Digital Libraries, and Individuals." *International Journal of Digital Curation* 1(1): 3-16. https://doi.org/10.2218/ijdc.v1i1.2

Berghmans, Stephane, Helena Cousijn, Gemma Deakin, Ingeborg Meijer, Adrian Mulligan, Andrew Plume, Sarah de Rijcke, Alex Rushforth, Clifford Tatum, Thed van Leeuwen, and Ludo Waltman. 2017. "Open Data: The Researcher Perspective." *Netherlands: Elsevier and Centre for Science and Technology Studies.* https://www.universiteitleiden.nl/en/research/research-output/social-and-behavioural-sciences/open-data-the-researcher-perspective

Center for Computational Science, (CCS). 2015. "LINCS Data Portal." http://lincsportal.ccs.miami.edu/dcic-portal

Center for Computational Science, (CCS). 2018. "Pulley Ridge Decision Support Resource." University of Miami. https://mesophotic.ccs.miami.edu

Die, David, Mahadev Bhat, Emily Starnes, Brett Pierce, and Timothy Norris. 2018. "Bio-economics: Historical Patterns of Fisheries and their Value: Evaluation of Future Policy Options." *University of Miami*. http://mesophotic.ccs.miami.edu/explore/module/bio-economics

Fecher, Benedikt, Sascha Friesike, and Marcel Hebing. 2015. "What Drives Academic Data Sharing?" *PLoS ONE* 10(2): e0118053. https://doi.org/10.1371/journal.pone.0118053

Fry, William. 1965. "Methods in Taxonomy." *Nature* 207: 245-246. https://doi.org/10.1038/207245a0

Giaretta, David. 2008. "Whitepaper: DCC Approach to Digital Curation - under Development." *Digital Curation Centre*.

Heidorn, P. Bryan. 2008. "Shedding Light on the Dark Data in the Long Tail of Science." *Library Trends* 57(2): 280-299. https://doi.org/10.1353/lib.0.0036

Holdren, John. 2013. “Memorandum for the Heads of Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research.” *United States Office of Science and Technology Policy*. https://petitions.obamawhitehouse.archives.gov/petition/require-free-access-over-internet-scientific-journal-articles-arising-taxpayer-funded

Iyengar, Shanto, and Douglas S. Massey. 2018. "Scientific communication in a post-truth society." *Proceedings of the National Academy of Sciences* 116(16): 7656-7661. https://doi.org/10.1073/pnas.1805868115

Johnston, L.R., J. Carlson, C. Hudson-Vitale, H. Imker, W. Kozlowski, R. Olendorf, and C. Stewart. 2018. "How Important is Data Curation? Gaps and Opportunities for Academic Libraries." *Journal of Librarianship and Scholarly Communication* 6(1): eP2198. https://doi.org/10.7710/2162-3309.2198

Mader, C., A. Koleti, C. Chung, H. Kucuk-McGinty, D. Vidovic, U. Vempati, S. Abeyruwan, D. Puram, N. Datar, S. Dhabe, R. Pissano, U. Visser, and S. Schurer. 2015. "LIFEwrx." http://life.ccs.miami.edu

NCCOS. n.d.a. "Understanding Coral Ecosystem Connectivity in the Gulf of Mexico from Pulley Ridge to the Florida Keys." Research Project Description. Accessed December 16, 2019. https://coastalscience.noaa.gov/project/coral-ecosystem-connectivity-gulf-florida-keys

NCCOS. n.d.b. "Understanding Coral Ecosystem Connectivity in the Gulf of Mexico from Pulley Ridge to the Florida Keys." Dataset Collections. Accessed December 16, 2019. https://products.coastalscience.noaa.gov/collections/regional/pulleyridge

Padilla, Thomas, Laurie Allen, Hannah Frost, Sarah Potvin, Elizabeth Russey Roke, and Stewart Varner. 2019. "Final Report --- Always Already Computational: Collections as Data (Version 1)." *Zenodo*. http://doi.org/10.5281/zenodo.3152935

Rittel, Horst W. J., and Melvin M. Webber. 1973. "Dilemmas in a General Theory of Planning." *Policy Sciences* 4(2): 155-169. https://doi.org/10.1007/BF01405730

Smith, R., V. Kourafalou, and A Valle-Levinson. *pending*. "Coral Ecosystem Connectivity from Pulley Ridge to the Florida Keys: Pulley Ridge, Dry Tortugas, and southwest Florida Shelf CODE/DAVIS and SVP Surface Drifting Buoy Trajectories from DATES." Edited by Atlantic Oceanographic and Meteorological Laboratory (AOML). Coral Gables: University of Miami.

Stuart, David, Grace Baynes, Iain Hrynaszkiewicz, Kate Allin, Dan Penny, Mithu Lucraft, and Mathias Astell. 2018. "Whitepaper: Practical challenges for researchers in data sharing." *figshare*. https://doi.org/10.6084/m9.figshare.5975011.v1

Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. 2011. "Data Sharing by Scientists: Practices and Perceptions." *PLoS ONE* 6(6): e21101. https://doi.org/10.1371/journal.pone.0021101

Tenopir, Carol, Elizabeth D. Dalton, Suzie Allard, Mike Frame, Ivanka Pjesivac, Ben Birch, Danielle Pollock, and Kristina Dorsett. 2015. "Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide." *PLoS ONE* 10(8): e0134826. https://doi.org/10.1371/journal.pone.0134826

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data* 3: 160018. https://doi.org/10.1038/sdata.2016.18