Cyberinfrastructure to Support Data Management

Rob Bochenek
Axiom Data Science
Anchorage, USA
rob@axiomdatascience.com

Chris Turner
Axiom Data Science
Anchorage, AK USA
chris@axiomdatascience.com

***Abstract* — Management of oceanographic data is particularly challenging due to the variety of protocols for data collection and analysis and the vast range of oceanographic variables studied. This paper describes the end-to-end cyber infrastructure developed to support stakeholders in the ocean science community throughout the data life cycle: from immediately after data collection through numerical analysis and synthesis, visualization, and decision making, to data publication and reuse. Our intent is to provide an overview of the system architecture and descriptions of system components.**

Keywords—data management, cyberinfrastructure, workflows, HPC, metadata

I. INTRODUCTION

The reasons to strive for better management of oceanographic data are many (e.g., reuse of old data for new research, reproducibility of results, preservation of the scientific record, etc., see [1], [2], [3]), though challenges and disincentives to effective, widespread data archiving, sharing, access, and reuse persist. Recent research describes a variety of barriers that scientists must overcome to make their data available for long-term preservation, public access, and reuse. Researchers report being concerned that their data will be misinterpreted due to its inherent complexity or poor quality control from potential reusers [4]; having insufficient time to curate or prepare their data for long-term data management (ibid); and lacking funding for managing their data beyond the timeframe of their research projects (ibid). Tenopir et al. [4] found almost two-thirds of scientists surveyed reporting lack of access to support staff dedicated to long-term data management. This is especially concerning given that data managers from the International Polar Year (IPY) found that effective management of the interdisciplinary data typical in ocean and coastal research required increased interpersonal interaction between the data producers and the technologists and curators responsible for providing long-term data archiving and access services as well as the opportunity for scientists to actively participate in data curation [5]. As a further barrier to such participation, scientists and the data management community lack shared, easy-to-use tools for data curation [4], [6], especially those that address the entire data lifecycle [7].

Managing oceanographic data is particularly challenging due to the variety protocols for data collection and analysis and the vast range of oceanographic variables studied. Data may be derived from automated real-time sensors, remote sensing satellites, observational platforms, field and/or cruise observations observations, model outputs, etc. Variables can range from large-scale ocean dynamics to micro-scale zooplankton counts. The resulting datasets are packaged and stored in advanced formats and describe a wide spectrum of scientific observations and metrics. Due to the complexity of the data, developing data management strategies to securely organize and disseminate information is also technically challenging. Distilling the underlying information into usable products for diverse user groups requires a cohesive, end-to-end approach in addition to a fundamental understanding of the needs and requirements of the dataset's creators, end users, and community stakeholders.

Over the last decade, Axiom Data Science (Axiom) has worked with state, federal, and private partners to develop the technologies and capabilities necessary to address many of the common challenges to ocean data management, reuse, and visualization, including securely storing and sharing data within research teams and larger research campaigns; providing tools for scientists to perform reproducible analytical workflows; publishing data with standards-compliant metadata; assimilating and visualizing data in ways that allow data of heterogeneous types and spatiotemporal granularities to be integrated, explored, and understood together; and efficiently accessing and analyzing high-volume data products, including model results and satellite imagery. By combining these technologies, we have created an end-to-end data management ecosystem that provides scientists with tools for meeting their data-related obligations and performing collaborative analyses using reproducible workflows. These tools also ensure that principal investigators, data managers, and program supervisors have a transparent view of project progress with respect to data collection, documentation, and publication tasks and deadlines, and they enhance the impact, reuse, and accessibility of ocean science research products by making them available to decision makers and other interested stakeholders alongside other observational, in situ, remote, and real-time data products from other ocean and coastal research and monitoring efforts.

The individual components of this ecosystem were developed from open-source technologies to be scalable and to embrace community standards and best practice recommendations for data and metadata management. These standards and technologies encompass, but are not limited to Jupyter Notebooks with support for the R and Python languages, the NetCDF Climate and Forecast (CF) Conventions, various Open Geospatial Consortium (OGC) standards for data interoperability, the ISO 19115-based suite of geospatial metadata standards, the Digital Object Identifier

(DOI) standard, the DataONE Member Node package, and various software implementations for data access and discoverability, including THREDDS, ERDDAP, Geoserver, OPeNDAP, and Sensor Observation Service (SOS). As a part of the DataONE Network, our data ecosystem is seamlessly integrated with more than 40 other archives that participate in the DataONE Network as member and replication nodes, providing users of our ecosystem with access to the nearly 100,000 datasets archived in other member nodes, and ensuring that datasets published from our data system are discoverable at other archives across the network.

II. Data System Approach

Informed by more than 10 years of experience providing data management services to coastal and ocean research and monitoring efforts, Axiom has developed a framework for managing oceanographic data (Fig. 1). This framework provides tools to help scientists manage data during the planning, collection, and analysis phases of their work; stores data in native and more useful, transformed formats; exposes data through interoperability systems; automates pathways for submitting data to national data centers for preservation and publication; and integrates several user interface tools to allow the data to be manually and programmatically discovered and explored by the broader community.

A. Data System Tier 1: Data, Models and Metadata

At the base of the data system framework are the datasets, metadata, and model outputs that provide the foundation for applications and user tools. These resources can be stored either in native formats or converted into spatially-enabled databases for storage and access. The decision to choose one method over the other is dictated by the requirements of the interoperability system that will be serving the data. Data that has a tabular or vector form (Shapefiles, databases, Excel spreadsheets, comma separated values (CSV) files, etc.) are saved in their native formats and converted into netCDF files when appropriate. After any reformatting is complete, the data are loaded into a PostgreSQL database and spatially indexed. When possible, GeoServer, an open-source geospatial data server, is then connected to the database and serves the data via WFS and WMS protocols. Imagery, raster data, and model results are stored on a file server in their native file formats. THREDDS and/or ncWMS are used to serve netCDF and HDF files, which may contain two, three, four, or higher-dimensional gridded datasets. GeoServer or other OGC-compliant mapping servers are used to serve GeoTIFF, ArcGrid, or other two-dimensional imagery or raster data.

B. Data System Tier 2: Interoperability Systems

Various interoperability servers (GeoServer, THREDDS, ncWMS, ERDDAP, OpenDAP, 52 North SOS, etc.) are implemented on top of source data to expose a powerful set of interfaces for other computing systems and humans to extract, query, and visualize the underlying source data. These systems provide many, redundant options for providing data to users in their preferred formats, in addition to providing the mechanisms for machine-to-machine data transfer to national data assembly and archive systems as required. These systems have been developed using the Java programming language and run within Tomcat servlet containers.

C. Data System Tier 3: Asset Catalog, Ontological Metadata, and Services

The asset catalogue provides a description of known internally- and externally-available data resources, access protocols for these resources (interoperability services, raw file download, etc.), and directives on how to ultimately utilize these data resources in applications. Because documentation and access methods vary widely between data sources, a system that catalogs data sources and reconciles these inconsistencies is essential for the data to be used in an efficient manner. In addition to managing information about data availability and access methods, the asset catalogue also contains ancillary data such as geographic locations, spatial and temporal resolutions, units, CF parameter(s), and information about the sources of the data.

D. Data System Tier 4: User Applications

Web services written in Java and Python connect to the asset catalogue and provide applications with access to the underlying descriptions of data assets and sources. Because the asset catalogue contains relationally-structured maps between data types, sources, and a controlled set of definitions, all front-end applications can connect users to vast arrays of data through simple but powerful interfaces. These interfaces include the following:

- public-facing catalogs of data assets that are updated automatically when new data is ingested into the system;
- a powerful, prioritized search interface that allows users to search by geography, time, access method, or words contained in metadata descriptions;
- a secure method to share project and dataset metadata and files with the public through the data catalogs; and
- interactive maps that allow users to explore spatial data as well as compare them to other, related datasets.

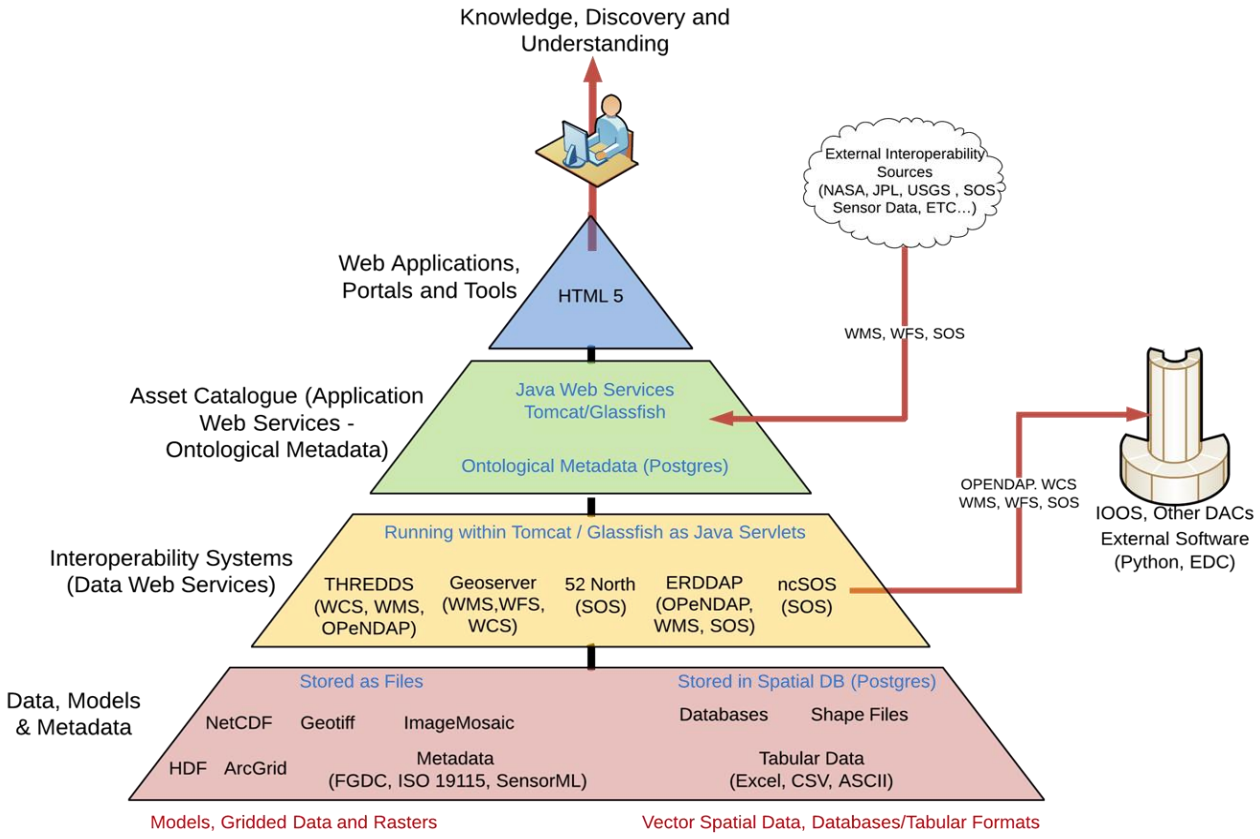

Fig. 1. A conceptual representation of the data system developed by Axiom that details the flow of data through logical technology tiers. The data itself is the foundation of the data system, with accessibility and value increased in each layer, and building towards being discovered and used to enable knowledge, discovery, and understanding.

III. Data System Components

This paper does not provide a comprehensive inventory of data system components. Instead, we describe below interesting and representative pieces of the larger system.

A. Physical Infrastructure

All physical infrastructure required to support Axiom's data system is located in a data center designed and maintained by Axiom staff. Resources in Axiom's data center include more than 2,500 processing cores arranged in a series of interconnected blade arrays, as well as slightly more than 1 petabyte of usable storage that includes multiple redundant backups. Compute nodes and storage nodes are connected over a low latency, converging network fabric (40 Gb/s Infiniband). GlusterFS is employed as a storage software abstraction layer that enables clients and storage servers to exploit data transfer over Remote Direct Memory Access (RDMA) protocols. This configuration enables data throughput from the storage clusters to the compute clusters to reach speeds greater than 160 Gb/s in high-concurrency situations. Axiom's Anchorage office also has a dedicated multi-braided 1 Gb/s high-speed internet connection for large file transfers between external data centers and for high-bandwidth demands of centralized web based applications. Axiom provides the following enterprise-level infrastructure capabilities:

1. *Security and Redundancy*

Axiom designed and maintains its own data center, collocated with the Pittock Internet Exchange in Portland, OR, part of the West Coast US internet backbone. There, the data center benefits from the low-latency, high-bandwidth internet connection, and network and power reliability. All data center resources are protected by several levels of onsite redundancy and backup, with offsite backup through Amazon Glacier. This design ensures that multiple redundant copies of data exist in addition to web application servers. Several layers of physical hardware (enterprise-level firewalls) and system monitoring software (NAGIOS) are also in place to provide hardened cyber security.

2. *Capacity and Performance*

High Performance Computing (HPC) has been a component of the Axiom technical strategy since early 2011. Axiom operates its own private cloud of compute and storage resources that data managers can provision to specific tasks and roles. The current numbers of processing cores and storage volumes are scalable to allow additional resources to be added as necessary. Axiom engineers have demonstrated that large GIS, model, and remote sensing datasets require HPC environments to be visualized and queried over web-

based interfaces. Because HPC is achieved through load balancing and parallelization, these types of systems also provide the added bonus of high availability and redundancy.

B. Real-time Sensor Stack

Axiom has developed cyberinfrastructure for managing and visualizing high-volume, heterogeneous, real-time observations from in situ devices. The real-time sensor stack is a specific instance within the general system described in the prior section of this paper, Data System Approach. Each week, millions of observations from more than 30,000 stations are ingested by Axiom and cached in fast, high-availability resources. Data formats and metadata are improved on import, with ingestion processes varying as needed for hundreds of data sources. Observations are added from the cache to netCDF-formatted archives of historical observations for each station. These historical sensor archives cover the entire time during which data from the station is ingested, plus any older observation values available from the data source. This component of the data system powers the Integrated Ocean Observing System's (IOOS) Environmental Sensor Map, described below and shown below on Fig. 2.

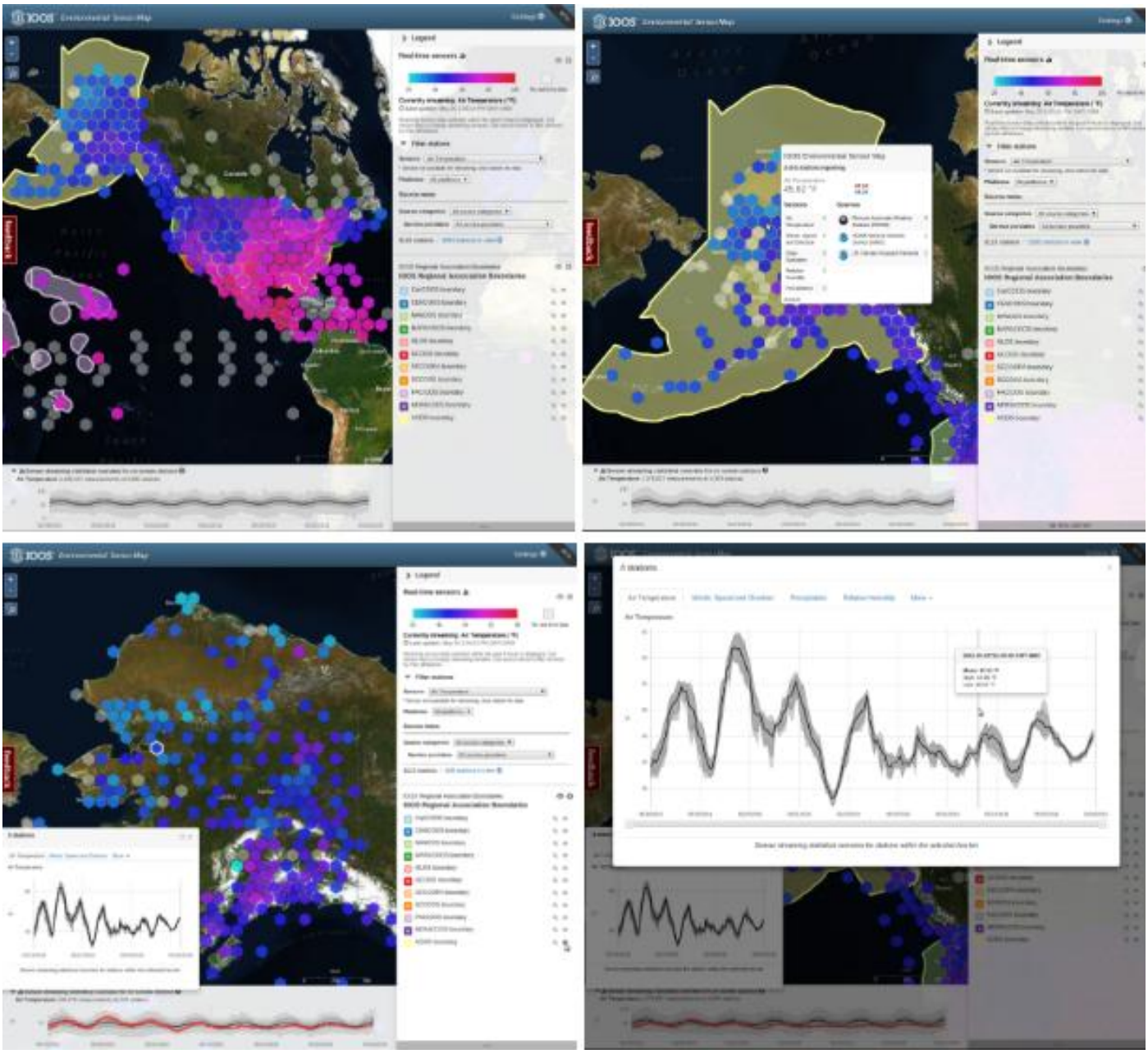

Fig. 2. Screen captures of The IOOS Environmental Sensor Map showing, from left to right, top to bottom: a hex-binned heat map (literally) of values from streaming air temperature sensors across the US and a statistical overview for the area; the Alaska region of the IOOS map, showing the AOOS region outlined in yellow and hex bins of streaming air temperature sensors map bounds; a synthetic time series generated from the air temperature sensors in the selected hex bin; an enlarged view of the synthetic time-series that allows users to explore narrower temporal subsets of the time-series, to mouse over the graphical summary for data values at each time instant in the series, and to switch between parameters observed by sensors in the binned area.

C. *Thematic Data Portals*

Data are most valuable when they are used, on their own or in synthesis with other datasets, to create new understanding and inform decision making. In order for that value to be realized, data must be accessible, understandable, and usable. Through a variety of thematically focused data portals, the data and services that make up the Axiom data system are exposed to the public. These portals serve as public-facing, easily-navigable and searchable sources for a variety of types of datasets, for metadata describing the datasets and their sources, and for data products and visualizations for exploring the data and the relationships between parameters across datasets and over space and time.

1. *AOOS Ocean Data Explorer (ODE)*

This data catalog and portal is the flagship data product of the Alaska Ocean Observing System (AOOS). The ODE provides a single access point for operational oceanographic and atmospheric models, satellite imagery, real-time sensor feeds, GIS datasets, and ground- and ship-based observations and measurements describing the biological, chemical, and physical characteristics of Alaska and its surrounding waters. The ODE provides access to all of AOOS's public data organized into more than 2,800 distinct data collections or modules, which are comprised of tens of thousands of distinct data files and layers. The portal allows users to integrate and visualize different types of data from many sources on an interactive map; add and remove layers, select from multiple base maps, and see changes in the selected layers over time with an interactive time slider.

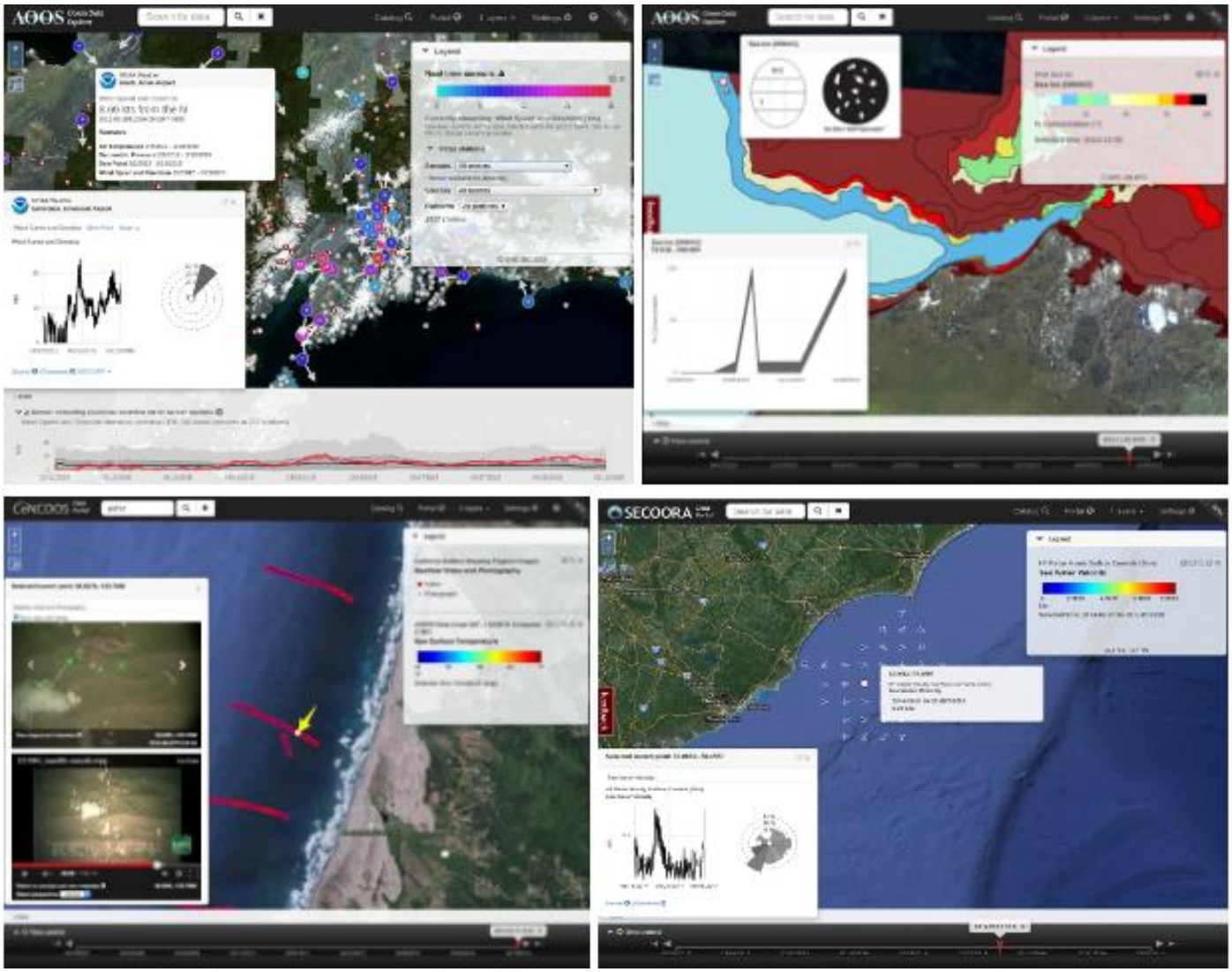

Figure 3. Screen captures from data portals. Showing, from left to right, top to bottom: real-time sensors in the AOOS region, showing currently streaming sensors, a wind rose with directions from the past seven days, and a statistical overview of the area (AOOS ODE); the Shell Ice and Weather Advisory Center's sea ice forecasts, visualizing dynamic sea ice vector layers produced by Shell for Arctic operations (AOOS ODE); sea surface temperature from the Advanced Very High Resolution Radiometer processed into a one month composite (CeNCOOS data portal); high frequency radar measurements of offshore surface currents in (SECOORA data portal).

The catalog and portal in the ODE provide access to metadata and project contacts, as well as web services to subset and download data in a variety of formats. The AOOS ODE[1] is the most varied and high-volume thematic portal Axiom has created; this is largely a function of AOOS's outsized role in ocean science coordination and management in Alaskan waters. Axiom has created similar regional portals for Central and Northern California Ocean Observing System (CeNCOOS) [2] and Southeast Coastal Ocean Observing Association (SECOORA)[3], both of which are shown with the AOOS ODE in Fig. 3, above.

2. IOOS Environmental Sensor Map[4]

The IOOS Sensor Map integrates millions of real-time observations each week into a single portal for data discovery and access. The most recent 30 days of data from all stations remain cached for quick access; older values are retrieved from the historic sensor data archive. By default, mapped stations are hex-binned to reduce on-screen clutter, with bins shaded to display a heat map of station geographic density. Incorporated tools allow the map to be filtered by sensor, to show a hex-binned heat map for the parameter measured by the selected sensor type; by platform, to display the geographic density by to platform type; and by station source. When zoomed in, the map displays individual station locations, which each station displaying basic station metadata on hover. Selecting any single station provides links back to the appropriate data source and dynamic graphs providing an overview of recent values from each sensor on the station. The sensor map also creates on-the-fly time series downloads with basic metadata, and provides an ERDDAP endpoint for specific, query-based time-series downloads with standardized metadata. The sensor map is shown in Fig. 2, below.

In addition to these large regional and national portals described above, the data system also powers several portals aimed at specific communities and one-off tools with particular uses. Examples include but are not limited to tools designed for creating on the fly visual summaries of billions of rows of seasonal and near real-time AIS ship traffic data; identifying the potential risk[5] and actual outbreak of harmful algal blooms[6]; distributing marine mammal health data[7]; locating in space and time all in-situ marine and coastal instruments and planned, underway, and completed research and monitoring efforts in order to assist with the coordination and planning of future efforts[8].

D. Research Workspace

The Research Workspace (the Workspace) is a web-based data management application designed and developed by Axiom specifically for use by scientists, research technicians, and project and program managers for storing, documenting, analyzing, and sharing data among members of scientific communities. Since the release in April of 2012 of the Ocean Workspace, the precursor to the Research Workspace, the user base has grown to more than 500 individuals from a number of large-scale scientific research programs funded by a variety of state, federal, and non-governmental organizations. Users have uploaded more than 18 terabytes of data spread across more than 800,000 files.

The Research Workspace provides users with an intuitive, web-based interface that allows scientists to create *projects* to represent particular scientific studies or research focuses or activities within a larger effort. Within each project, users may create topical groupings of data in folders and upload data and add contextual resources (e.g., documents, images, and any other type of digital resource) to their project by simply dragging and dropping files from their desktop into their web-browser. ISO 19115-2 metadata can be generated for both projects and their individual constituent datasets. Users of the Workspace are organized into research campaigns, and everyone within a campaign can view the projects, folders, and files shared with the campaign by other members. This allows preliminary results and interpretations to be shared by geographically- or scientifically-diverse individuals working together on a project or program before the data are shared with the public. It also gives program leads and other stakeholders a transparent and front-row view of how users have structured and described projects, and how their programs are progressing through time. The Workspace has the following capabilities:

1. Secure User Profiles

Users of the Workspace have a password-protected user profile that is associated with each of their projects. A user may associate their profile with the profiles of organizations to which they belong, and may associate their project(s) with one or more research campaigns. The interface allows users to navigate between projects, organizations, and campaigns in which they are involved through a menu or using integrated search tools. Transfer of data and information occurs over Secure Socket Layer (SSL) encryption for all interactions with the Workspace. The Workspace supports authentication through Google accounts, so if users are already logged into their Google account (Gmail, Google Docs, etc.), they can use the Workspace without creating a separate username and password.

[1] https://portal.aoos.org
[2] http://data.cencoos.org/
[3] http://portal.secoora.org/
[4] https://sensors.ioos.us/
[5] http://www.aoos.org/k-bay-hab/
[6] http://dev.axiomdatascience.com/?portal_id=97#map
[7] https://goo.gl/egt49D
[8] http://portal.aoos.org/research-assets

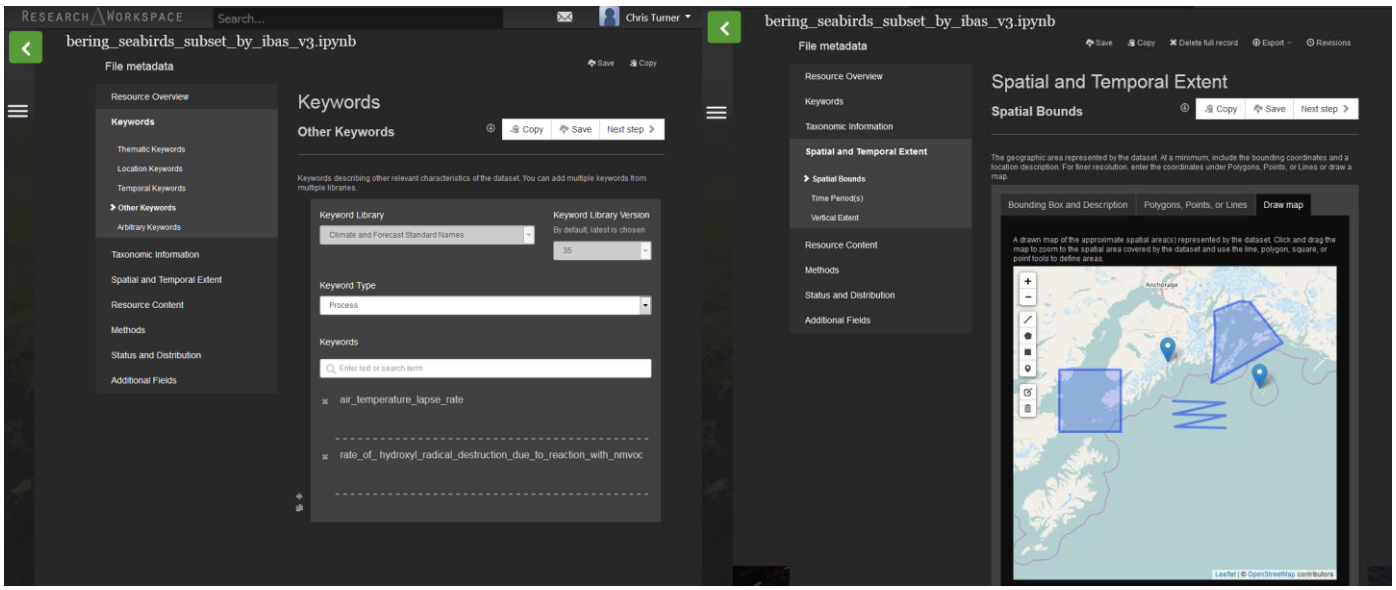

Fig. 4. Two screen captures of the Research Workspace's integrated metadata editor, showing from left to right: the tool for entering keywords which allows the user to select from several controlled vocabularies and from specific versions of those vocabularies; and an integrated tool for drawing bounding boxes and polygons, line strings, or points to represent the spatial extent of the project or dataset described by the metadata.

2. Metadata Authoring

Metadata fields in the Workspace's integrated metadata editor come from the ISO 19115 suite of standards for geospatial metadata, and the FGDC endorsed successor to the CSDGM, extended to describe taxonomic classification for biological datasets. To facilitate taxonomic description, Axiom developed a tool that allows users to search the ~625,000 taxonomic entities of the Integrated Taxonomic Information System (ITIS) in order to rapidly add species information to metadata. The editor also contains tools for copying content between records and for designating template records to be used within projects, organizations, or campaigns. Because the Workspace is a cloud-based service, researchers can move between computers during the metadata generation process in addition to allowing team members and administrators to simultaneously review and edit metadata in real time. Metadata can be exporting as XML files compliant with ISO 19139, the xml schema implementation for 19115-2 and 19110. Metadata authoring interfaces are shown in Fig. 4, above.

3. Advanced and Secure File Management

A core functionality of the Workspace is the ability to securely manage and share project-level digital resources in real time with version control among researchers and study teams. Users of the Workspace are provided with tools to bulk upload files, organize those documents into folders or collections, create projects with predefined and user-created context tags, limit access to their projects, and control read and write permissions on files within projects. The Workspace also provides a simple form of version control: when a user re-uploads a file of the same name, the most current version of the file is displayed, but perpetual access is provided to past versions as well.

4. Reproducible Analytical Workflows

Reproducibility is fundamental to community trust in scientific results, to public trust in the scientific process, and to transparency in the use of public resources in the service of science and science-informed resource management [8]. For scientific processes with computational components to be reproducible, the many scripts, commands, runs, and results used for data processing and analysis should be made publicly available [9]. To meet this need, the Workspace instance in the Workspace includes kernels for several versions of the Python and R languages, all of which can be interpreted in the same notebook. By moving data analysis to the web, workflows can be more easily and broadly shared and examined. By making the code itself available with embedded, human readable explanation, the workflows and their results are made more understandable to those who would reuse them or apply them towards new applications, and their inputs and dependencies are made explicit.

Building the Jupyter Notebooks environment into the Workspace itself allows a user to create notebooks that operate on any data in the Workspace or the entire data system that the user has permission to access. By default, this includes hundreds of terabytes of model, satellite, and real-time sensor data archives, as well as many other geospatial products and any resources in the user's Workspace projects. When executed, these notebooks run on dedicated HPC resources in the same data center as the storage clusters that are home to all of the data system assets. This collocation of resources significantly reduces data transfer times and accelerates the analysis of very large environmental and oceanographic datasets. Fig. 5, below, shows a Jupyter Notebook in the Workspace.

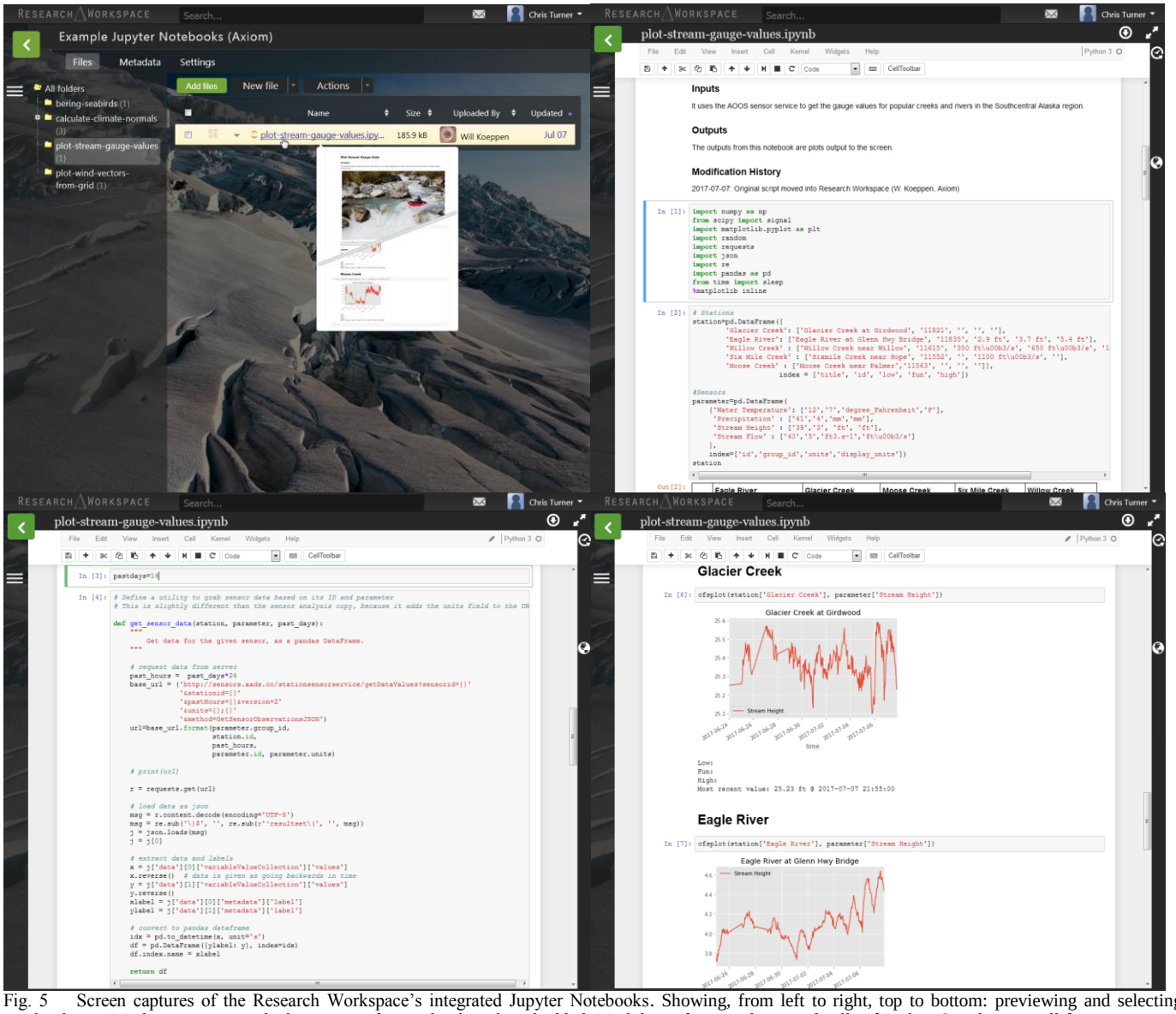

Fig. 5 Screen captures of the Research Workspace's integrated Jupyter Notebooks. Showing, from left to right, top to bottom: previewing and selecting a notebook in a Workspace project; the beginning of a notebook with embedded, Markdown-formatted text and cells of Python 3 code; one cell for user input and long cell of Python code; and the output graphs of stream gauge readings over time, with river level classifications when available.

5. *Publishing and Preservation*

Long-term preservation of data is essential for dependably making the results of a scientific project available for reuse beyond the life of the project or the career of the scientist that generated it. Though the Workspace provides security and redundancy for all its content and can expose datasets through Axiom's data portals, it is not designed to serve as an environment for long-term preservation or publication. To achieve this, the Workspace has been seamlessly connected to an external data repository maintained by Axiom staff. This repository is an instance of the DataONE Generic Member Node[9], a preservation-oriented repository that provides byte-level persistence of objects, accepts and exposes any metadata format, and is connected to the DataONE Network. As a DataCite[10] member organization, Axiom is able to reserve and assign DOIs to all archived objects in the Research Workspace DataONE Member Node. As a member of the DataONE Network, all content in the Research Workspace Member Node is replicated at several other geographically-distributed data archives across the network. This distributed redundancy ensures the security of the data against local failures and its availability in perpetuity independent of the future of Axiom or any single replication center.

6. *Specific Technical Components*

a) Database systems

PostgreSQL 9 is used for storage of tabular and relational data representations, and is extended with PostGIS for spatial data. All data uploaded to the Workspace are

[9] https://dataone.org/software-tools/generic-member-node

[10] https://datacite.org/

replicated across multiple database servers to provide redundancy and ensure high availability.

b) Object storage and data representation

MongoDB is used as a persistent NoSQL storage and query system for file objects, tabular data (flat structures), and hierarchically structured data (generally XML). MongoDB allows horizontal scaling through sharding across physical devices and provides redundancy and high availability through replication. The MongoDB instance consists of a three-node cluster, and each node maintains a complete replicate of the others. Data within each node are further redundant by virtue of RAIDed disk arrays.

c) Web tier

The web services used by the Workspace are developed using Java and integrated into a web application framework called Play!, which provides a stateless architecture for Java and Scala development. The RESTful, stateless design allows services to be scaled across application nodes for load balancing, redundancy and horizontal scalability. The framework is also used to provide real-time notifications browser clients to enable collaboration amongst users.

d) Caching and pub/sub

Redis is used as an intermediary between the web and data tiers. It also serves as our pub/sub interface for managing communications between web tier nodes and serving real-time connections to browser clients in a scalable manner.

e) User interface

The user interface of the Workspace is composed of several JavaScript and HTML5 libraries and integrates with server-side modules wrapped into the Play! framework. The frontend uses a client-side MVC architecture in Backbone.js that synchronizes with its backend equivalent to provide users with a more responsive experience than is typically found in many web applications.

IV. WRAP UP

The Axiom data system is a centralized stack of technologies designed around the need to accommodate the diversity of data management needs and data types used in the ocean sciences. The Research Workspace provides a web-based environment for sharing, documenting, analyzing, and publishing data for planned and in-progress efforts. Data in the Workspace and from completed projects or external sources are stored on high volume, redundant storage clusters collocated with HPC processing resources and ultra-low latency, high-bandwidth network access and reliable hydroelectric backup power. Standardized and community-developed interoperability systems connect all stored data to the asset catalog, a relational, pseudo-ontological metadata store describing data and its sources. These and other interoperability services provide programmatic access to the assets in the data system to external data assembly centers and other technical data users. Axiom's data portals and other web applications access data and metadata through the asset catalog.

Many research groups, organizations, Integrated Ecosystem Monitoring Programs, and individual scientists are currently leveraging some or all of the components of our cyberinfrastructure for their data analysis, management, publication, or visualization needs and in pursuit of the their larger research, monitoring, or decision-support goals. Some of our current partners include the IOOS Office, AOOS, CeNCOOS, SECOORA, the Marine Biodiversity Observation Network, the North Pacific Research Board, the Exxon Valdez Oil Spill Trustee Council, the Bureau of Ocean Energy Management, the National Oceanographic Atmospheric Agency, the Long Term Ecological Research Network, and the Department of Homeland Security, among others.

ACKNOWLEDGMENT

The authors would like to thank all of our partners that use and have provided valuable feedback on our data management strategy and system, and all of the staff at Axiom Data Science.

REFERENCES

[1] B. Fecher, S. Friesike, and M. Hebing, "What Drives Academic Data Sharing?" *PLoS ONE*, vol. 10, no. 2, e0118053, Feb, 2015. doi:10.1371/journal.pone.0118053

[2] T. Hey, S. Tansley, and K. Tolle, Eds. *The fourth paradigm: Data-intensive scientific discovery* .Redmond, WA: Microsoft Research, 2009.

[3] P. Soranno, K.S. Cheruvelil, K.C. Elliott, and G.M. Montgomery. "It's good to share: Why environmental scientists' ethics are out of date", *BioScience*, vol. 65, no. 1, pp. 69-73, Oc 2014. doi:10.1093/biosci/biu169

[4] Tenopir C, et al. "Data sharing by scientists: Practices and perceptions", *PLoS ONE*, vol. 6, no. 6, e21101, June 2011. doi:10.1371/journal.pone.0021101

[5] Parsons M, et al. "A conceptual framework for managing very diverse data for complex, interdisciplinary science", *Journal of Information Science*, vol. 37, no. 6, pp. 559-569, Oct 2011. doi:10.1177/0165551511412705

[6] P.N. Edwards, M.S. Mayernik, A.L. Batcheller, G.C. Bowker, and C.L. Borgman, "Science friction: Data, metadata, and collaboration", *Social Studies of Science*, vol. 41, no. 5, pp. 667-690, Apr 2011. doi:10.1177/0306312711413314

[7] S. Abrams, et al., "DataShare: Empowering researcher data curation", *International Journal of Digital Curation*, vol 9, no. 1, pp. 110-118, Feb 2014. doi:10.2218/ijdc.v9i1.305

[8] L.C. Burgess, et al., "The Alan Turing Institute Symposium on Reproducibility for Data-Intensive Research – Final Report", St. Hugh's College, Oxford, April 2016.

[9] G.K. Sadve, A. Nekrutenk, J. Taylor, and E. Hovig, "Ten Simple Rules for Reproducible computational research", PLoS Comput Biol, 9(10): e1003285. doi:10.1371/journal.pcbi.1003285