# The Big Data Lifecycle in Open Ecoinformatics: Curation, Analysis, and Sharing

A Dissertation
Presented in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
with a
Major in Natural Resources
in the
College of Graduate Studies
University of Idaho
by
Edward Flathers

Approved by:
Major Professor: Paul E. Gessler, Ph.D.
Committee Members: Jeremy Kenyon, MLS; Xiaogang Ma, Ph.D.; Lucas Sheneman, Ph.D.
Department Administrator: P. Charles Goebel, Ph.D.

May 2022

# Abstract

Research data go through a cyclical process from the point of their conception during project planning, through experimental design and sample design, data collection, organization, analysis, storage, curation, and, ideally, re-use. Historically, not all steps in the lifecycle have been given the same level of attention.

Much of the data researchers have collected have become "dark data," often recorded on paper and, once the project has concluded, consigned to a file cabinet, never to be seen again. There is a reproducibility crisis in the sciences that is being slowly revealed to have quietly spread across many disciplines, casting doubt on the veracity of some published results. Even when methods are transparent and data published, we face challenges agreeing exactly what the rules are for sharing research data with each other.

Chapter 1 provides an introduction and background information about Big Data, the data lifecycle, the FAIR Data Principles, and concepts surrounding open science. Together, these topics provide a foundation and motivation for the material in the remaining chapters.

Chapter 2 applies the concept of service-oriented architecture from computer sciences to the task of designing an OAIS (Open Archival Information System) data repository. Such repositories are used to store, curate, and manage research data, and to provide visibility and access to research data that help to enable re-use.

Chapter 3 provides an example of using the concepts of open science to produce research products using transparent methods that are clearly reproducible. While generating a model predicting levels of organic carbon found in soil in the Northwestern United States, the key to ensuring that results are reproducible is to publish all research data and computer code used in analysis and preparation of those results.

Chapter 4 addresses the issue of how we express and agree upon common rules for data sharing. As data sharing becomes less personal, more distributed, and potentially more automated, we need formal ways of expressing sharing agreements. Furthermore, these agreements must be easily readable by both humans and machines to be effective.

Chapter 5 provides some concluding remarks and considers the material of the earlier chapters in the context of contemporary challenges accompanying the era of Big Data.

# Acknowledgments

This work would not have been possible without the generous support of my committee members, Jeremy Kenyon, Marshall Ma, Luke Sheneman, and Paul Gessler. As my major professor, Paul has always been encouraging, patient, and enthusiastic as we worked out the details of these chapters. I would like to thank Sara Nelson, Bruce Godfrey, and Greg Gollberg for their recommendations, support, and advice about graduate school. Luigi Boschetti, Lily Wai, and Gail Eckwright all contributed to putting me on the path toward the spatial sciences. My friend Manuel Welhan has been a steadfast presence and sounding board throughout this project. I would also like to recognize the faculty and staff of the Department of Forest, Rangeland, and Fire Sciences for fostering an academic environment that helped me achieve my goals.

This material is based upon work that is supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture, under award number 2011-68002-30191; and by the Federation of Earth Science Information Partners.

# Dedication

Thank you to Jamie Flathers, Ronald Flathers, Barbara Flathers, Keith McCully, Sydney Reed, and Douglas Storkovich, without whose kind support these pages would not have been possible.

# Table of Contents

# List of Tables

# List of Figures

# Statement of Contribution

Statements of work using the CRediT author statement system:

Chapter 1: Introduction

[no coauthors]

Chapter 2: A service-based framework for the OAIS model for earth science data management

E. Flathers: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing

J. Kenyon: Writing - Review & Editing

P.E. Gessler: Writing - Review & Editing

Chapter 3: Building an Open Science Framework to Model Soil Organic Carbon

E. Flathers: Conceptualization, Methodology, Software, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing

P.E. Gessler: Writing - Review & Editing

Chapter 4: Methods for Expressing Machine-Readable License Information in Geospatial Metadata

E. Flathers: Conceptualization, Methodology, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing

J. Kenyon: Writing - Review & Editing

L. Sheneman: Writing - Review & Editing

M. Ma: Writing - Review & Editing

P.E. Gessler: Writing - Review & Editing

Chapter 5: Conclusion

[no coauthors]

# Chapter 1: An Introduction of Terminology

## Introduction

This chapter serves to introduce some of the terms used in the following chapters, and to provide a framework for understanding the chapters as parts of a whole: as steps in the data lifecycle. Already, an undefined term has appeared (the data lifecycle), but that will be addressed soon.

The first order of business is to define the term “Big Data,” which appears first in the title of this dissertation. Big Data is a concept that evades definition, but is useful as a way to think about data that require nontraditional or unconventional approaches to processes like storage, transmission, and analysis. The concept of Big Data is more about approaches to working with data than it is about any special “bigness” of data, themselves.

This dissertation is about the Big Data lifecycle, so the next important concept is the lifecycle of data. There are many models and explicit specifications of the data lifecycle, and one—the DDI (Data Documentation Initiative)—is here chosen for purposes of discussion; others might be chosen and discussed with similar vigor.

The title of this dissertation also contains the terms curation, analysis, and sharing. Data curation is a practice that combines library sciences with data sciences, delving into questions about the value and reusability of data and whether or not, or for how long, data should be preserved in case they could be useful in the future. Data analysis is straightforward, in theory, but is complicated by the replicability crisis that threatens trust in the reliability of results of scientific research not just in the fields in which it has already been identified, but in all fields of science. The concept of sharing science is age-old: the ideal purpose of the scientific journal is for scientists to share their experiences in such a way that others can build upon them. In this time of automation, the traditional methods of sharing science data are no longer adequate, and we need to consider more formal (and automatable) methods of doing so.

Therefore, this chapter will treat the concepts of Big Data and its definition; the data lifecycle and what it means in a data ecosystem in which data persist beyond their generating project; a set of rules governing the sharing of data called the FAIR data principles; and the concept of open science, which is an approach to scientific practice and publication that endorses and encourages a kind of radical publication – the exposition of not just the results and  products of scientific endeavor, as has been traditionally done in the journal article, but also the intermediate, possibly imperfect processing steps that lead to the final products that appear in the article – and the possibility that data collected for one purpose or project may eventually be reused for another project, perhaps in an entirely different context.

## Big Data

The origins of the term “Big Data” are not entirely clear (Diebold, 2012), and the term has resisted any standard definition since its introduction into, and common use in, the academic literature (Gandomi and Haider, 2015). However, a common thread tends to occur in definitions: the concept of the three (or more) Vs of Big Data. Originally enumerated as volume, variety, and velocity (Laney, 2001), and subsequently extended to include many more, the Vs of Big Data describe the characteristics that make data “Big.” There exists no canonical list of official Vs of Big Data, and the original three are adequate to this discussion.

Volume represents the measure of data size in bytes, which when large can result in data that resist traditional modes of storage and analysis. Variety represents heterogeneity of data, which can challenge systems such as relational models that tend toward narrowly defined data types and static organization. Velocity represents the speed at which data are generated or received; high-velocity data may be difficult to capture, for example, if they saturate the bandwidth of the communication medium.

Note that no specific numbers, such as a number of gigabytes of volume, are included here as minimums for data to qualify as Big. No agreement exists as to these minimum qualifications, and due to the continual progress of technology to increase storage capacity, processing power, and telecommunications speed, these numbers would necessarily be moving targets. Rather than using such figures to classify Big Data, it is more helpful to consider the techniques used to work with the data. Data might be Big if they are so large as to prevent storage on a single volume of disk, or if they require the implementation of new database designs or techniques to organize, or if their transmission over a line takes a prohibitive amount of time. Big Data is more about using novel approaches to work with data that push the boundaries of the computing environment than about specific numbers.

In this document, we use the concept of Big Data as a motivation for building general and robust repositories for research data, as an example of combining heterogeneous data from diverse sources, and in terms of sharing data in agreed-upon and automated ways.

## The Data Lifecycle

The concept of the data lifecycle serves as the framework for organizing the topics of the following chapters. The data lifecycle can be developed and envisioned at various levels of complexity and granularity, and for the purposes of this discussion can be restricted to a relatively sparse set of stages. Using the Data Documentation Initiative's (DDI) data lifecycle model (Figure 1.1), the first step in the lifecycle is the "study concept" at which the data do not yet exist, but where care should be taken to plan for their future within the research project and within the data lifecycle itself (Vardigan et al., 2008). Data are then collected and processed, which steps may conclude their relevance to the particular study. In order to support further activities involving data re-use, archiving, distribution, and discovery, steps are taken to store the data within a repository in which they will be findable and accessible. The final step in the DDI lifecycle model is "data analysis," where data are considered for re-use, and potentially repurposed back into the lifecycle as part of another project.

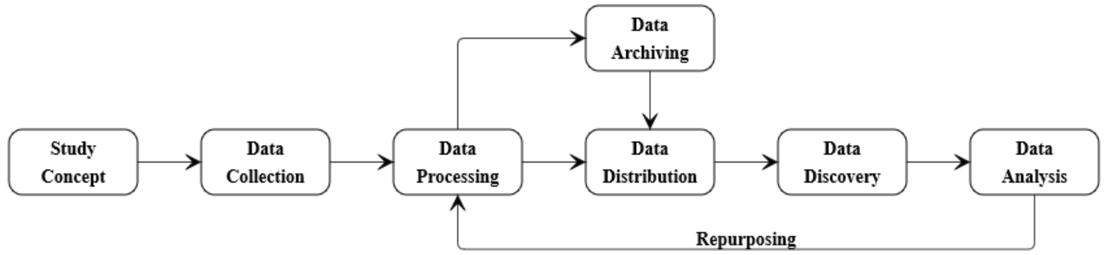

Figure 1.1 The DDI 3.0 data lifecycle

Chapter 2 of this document describes the implementation of a repository model that involves data archiving, distribution, and discovery. Chapter 3 describes a project that makes use of repurposed data, covering steps from data analysis back to data processing, and then continuing into the repository steps of the model. Chapter 4 involves establishing sharing agreements for data, which are involved in the repository steps: in the archival stage, a sharing agreement must be applied; in the discovery stage, the sharing agreement must be advertised, and in the distribution stage, the sharing agreement must accompany the data.

## The FAIR Data Principles

The FAIR Principles for scientific data management and stewardship were developed to improve the findability, accessibility, interoperability, and reusability of science data (Wilkinson et al., 2016). These four concepts occur roughly in chronological order during the process of reusing science data.

The first necessary step is to find or identify the existence of some data to be reused. Accessibility of data refers to the actions that must be taken to retrieve the data from wherever they are stored. Interoperability in this case describes how the data can be used, indicating, for example, whether particular software is needed to work with the data. Finally, reusability of data is the goal of actually incorporating the data into a new project.

The four concepts of FAIR data are implemented through a combination of data, metadata, and infrastructure. The data are obviously a necessary component of any data sharing or re-use goal. Metadata can be a powerful tool for enabling re-use. Science data should ideally always be accompanied by some form of metadata—that is, data describing the data. Metadata come in a variety of standards and formats, as well as a wide range of completeness and quality. Chapter 4 of this document is primarily focused on spatial data, and specific ways that we can enable data reuse by expressing rules for data sharing within metadata records in ways that can be easily read and interpreted by both humans and software programs. The third component of FAIR data, infrastructure, is also treated in Chapter 2 in detail. Finding and accessing data are made much easier when infrastructure such as data repositories exist, and when those repositories function according to standards for the preservation, distribution, and discovery of data.

## Open Science

Open Science is another term that has tended toward more frequent use in the academic literature despite having no clearly agreed-upon definition. Vicente-Saez and Martinez-Fuentes (2018) performed a systematic literature review to attempt a definition of open science, and concluded that "Open Science is the transparent and accessible knowledge that is shared and developed through collaborative networks."

Three key terms in this definition are "transparent," "accessible", and "collaborative networks." Transparency is a key issue in that science data analysis today is often performed using software and even data that may not be made public, leading to uncertainty as to the veracity of results—the heart of the reproducibility crisis in the sciences (Baker 2016, Gezelter 2015, McNutt 2012). An approach to transparency, advanced in Chapter 3, is the publication of all analytical objects of a study, including software code and data. Accessibility is, of course, one of the components of the FAIR Principles, and can be addressed through providing access to science data and code through repository systems, as addressed in Chapter 2. Collaborative networks require some infrastructure systems such as repositories to function, but they also require common sets of rules—the sharing agreements discussed in Chapter 4.

A recently-introduced tool, the Open Science Framework (OSF) "promotes open, centralized workflows by enabling capture of different aspects and products of the research lifecycle, including developing a research idea, designing a study, storing and analyzing collected data, and writing and publishing reports or papers" (Foster and Deardorff, 2017). The stages of the data lifecycle are clear in this description of the OSF tool, which functional as a working repository for projects and makes public the specifics of each step in the process.

Regardless of the specific definition used, the general goal of Open Science is transparency, or making public as much of the research process and product as is possible and practical. There are situations in which data may not be publishable due to laws or ethical constraints regarding personal privacy, official secrecy, intellectual property issues, and other concerns. Where these obstacles to publication occur, researchers are still encouraged to publish what is possible using anonymization techniques or careful omission of sensitive data. These restrictions do not usually apply to scientific software code, so publication is encouraged in order to ensure transparency. Computer code that is not published cannot be verified to accomplish the task it has been claimed to accomplish, eliminating transparency in the methods of research. Chapter 3 suggests a model for publication that includes publishing the exact code used to perform the analysis and to produce any data or figures that are used in the related paper.

## Conclusion

In the chapters to follow, these foundational concepts of Big Data, the data lifecycle, the FAIR data principles, and Open Science will be explored through applications that demonstrate and reinforce their utility in modern sciences. Chapter 2, A Service-Based Framework for the OAIS Model for Earth Science Data Management, describes a model for research data repository design based upon a modular, service-oriented architecture. Chapter 3, Building an Open Science Framework to Model Soil Organic Carbon, covers the development an open science framework for modeling organic carbon in soil in an area of the northwestern United States. Chapter 4, Methods for Expressing Machine-Readable License Information in Geospatial Metadata, explains how we can explicitly document the sharing agreements we choose to apply to research products and make those agreements accessible to both human and software consumers. Each chapter is a window into the Big Data lifecycle in open ecoinformatics. In addition, there is a common ideology behind these terms and concepts. The era of Big Data provides an opportunity to advance the practice of science through agreement upon standard, sharing, and transparency. The implicit underlying assumption is that the scientific community agrees that these are desirable qualities of scientific endeavor. These chapters will contribute to concepts surrounding open science processes, and a brief conclusion in Chapter 5 treats the necessary imposition of ideological intent when implementing open science methods, as well as some thoughts toward the future of science as data become more open and connected.

## References

Baker, M.  2016.  Is there a reproducibility crisis?  Nature 533.7604: 452–454.  doi:10.1038/533452a

Gandomi, A. and H. Murtaza. 2015. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35(2):137–144. doi: 10.1016/j.ijinfomgt.2014.10.007

Vardigan, M., Heus, P., and Thomas, W. 2008. Data Documentation Initiative: Toward a Standard for the Social Sciences. International Journal of Digital Curation 1(3):107–113 .

Diebold, F.X. 2012. A personal perspective on the origin(s) and development of "big data": The phenomenon, the term, and the discipline, second version (No. 13-003). Penn Institute for Economic Research, Department of Economics, University of Pennsylvania. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2202843. Accessed 27 February 2022.

Foster, E.D., and Deardorff, A. 2017. Open science framework (OSF). Journal of the Medical Library Association 105(2):203–206.

Gezelter, J.D.  2015.  Open source and open data should be standard practices.  The journal of physical chemistry letters 6.7:1168–1169.  doi:10.1021/acs.jpclett.5b00285

Laney, D. 2001. 3-D data management: Controlling data volume, velocity and variety [White paper]. META group research note.

McNutt, M.  2012.  Reproducibility.  Science 343 (6168), 229. doi:10.1126/science.1250475

Vicente-Saez, R., and Martinez-Fuentes, C. 2018. Open Science now: A systematic literature review for an integrated definition. Journal of Business Research 88:428-436. doi: 10.1016/j.jbusres.2017.12.043

Wilkinson, M. D., et  al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018. doi: 10.1038/sdata.2016.18

# Chapter 2: A Service-Based Framework for the OAIS Model for Earth Science Data Management

**Flathers, E., Kenyon, J. & Gessler, P.E. 2017. A Service-Based Framework for the OAIS Model for Earth Science Data Management. Earth Science Informatics, 10(3), 383-393. doi: 10.1007/s12145-017-0297-3**

## Introduction

Responsibility for managing data created in a laboratory or via field work has traditionally been held by researchers. Over time, this has led to a great diversity of scientific data management practices differing in thoroughness of documentation, application of technology, and preservation of data (Tenopir et al. 2011). As our capacity to collect data increases with the proliferation of sensor networks and new instruments and simulation methods, we face a “data deluge” that easily overwhelms many of our traditional data management efforts (Hey and Trefethen 2003). A significant component of the data deluge, and one that generates a growing level of funding opportunities in data management research, is so-called “big data” (Haendel et al. 2012).

Big data is a term used to describe data that are usually characterized by the “three Vs” of volume, velocity, or variety (Zikopoulos et al. 2011). Data of high volume are large in size and can require large storage and network bandwidth resources to manage. For example, in 2012, the National Institutes of Health’s 1000 Genomes Project exposed more than 260 terabytes of genetic data in more than 250,000 files (Clarke et al. 2012). Data of high velocity come into management systems, often from sensor systems, very quickly. The ATLAS Detector at the Large Hadron Collider creates 40,000,000 events per second, and filters out all but 200 per second, leaving them with a data recording rate of 320 megabytes per second (Haeberli et al. 2004). Data of high variety have inherent heterogeneity that can make them difficult to collate and compare. Take, for example, one minute’s worth of social media postings—2.5 million Facebook posts; 300,000 Tweets; 220,000 Instagram photos; 72 hours of YouTube videos— together, they tell a story about social media users, but each type of content requires a different set of tools for analysis (Gunelius 2014). Additional characteristics of big data have been identified in industry, but these three suffice to describe the challenges posed by big data in this paper.

One of the principal new challenges introduced by big data is data storage and curation (Hilbert and López 2011). As the information technology infrastructure needed to support research data grows in complexity and cost, the tasks of procurement and management can grow beyond the scope of the typical research project. Domain-specific and institutional data repositories have emerged to take advantage of economies of scale and provide standards-based methods for data storage and curation. To illustrate, over the past four years, the Registry of Research Data Repositories (re3data.org) has compiled a steadily growing registry of more than 1,500 research data repositories in more than 60 countries.

The definition and design of repositories has been developing in parallel to the emergence of the repositories, themselves. The Consultative Committee for Space Data Systems (CCSDS), in 2002, released their first version of a Reference Model for an Open Archival Information System (OAIS). In 2003, the model was adopted by the International Standards Organization (ISO) as ISO 14721:2003. The CCSDS document was updated in 2012 with additional focus on verifying the authenticity of data and developing concepts of access rights and a security model. The OAIS model is a good fit for research data repositories, having been designed as a framework to support data collections without regard to data types, storage formats, access methods, or other specific implementation details. Among other agencies, the Library of Congress, NASA, the ESA, and the USGS apply the OAIS model for science data management.

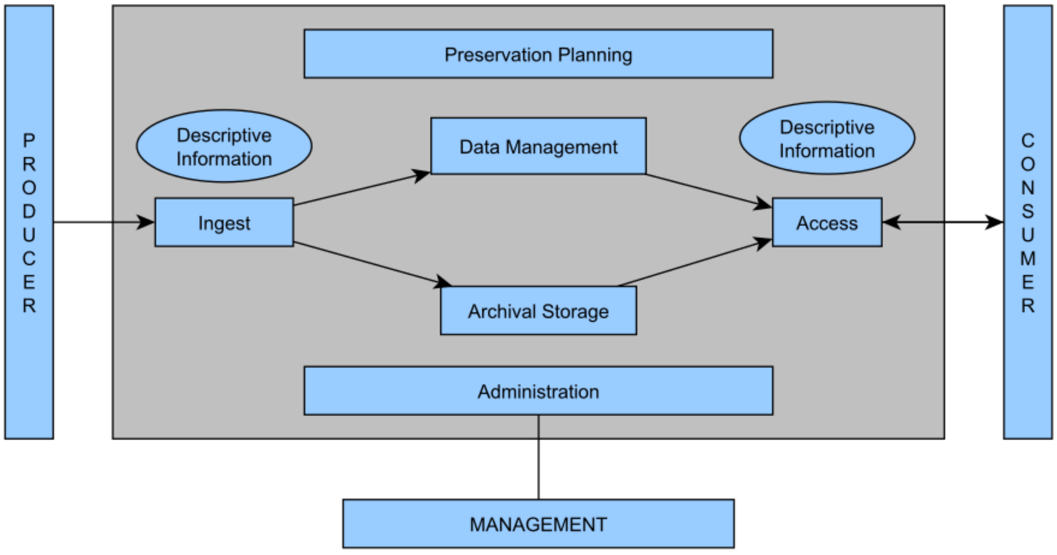

Figure 2.1 The block diagram of the OAIS model

The OAIS model involves bundling data and metadata into an Information Package that enables the basic functions of the repository: “ingestion, preservation, and dissemination of archived materials” (LaVoie 2014). There are four types of content contained within or associated with the Information Package: Content Information (CI), Preservation Description Information (PDI), Packaging Information, and a Package Description (PD).

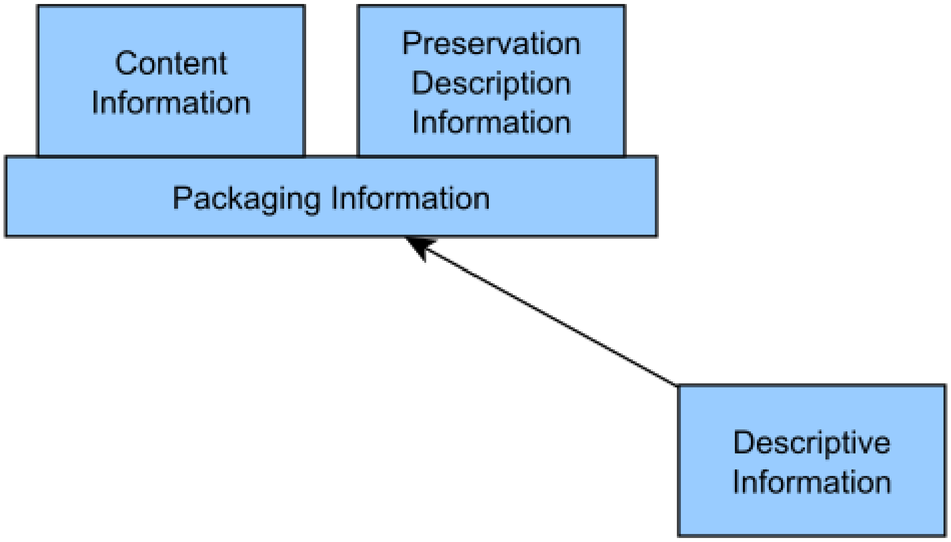

Figure 2.2 Information package concepts and relationships (after CCSDS 2012 Figure 2-3)

The CI is made up of a Content Data Object—the data content of the package, and Representation Information (RI)—the metadata associated with the data content. Data content is often stored in a format compatible with the software used to record it, such as Microsoft Excel, ArcGIS, and other general or specialized programs. The metadata stored in the CI describes the data: data identification, contact information, collection methods, accuracy assessments, and others. In the Earth sciences context, metadata are often stored in standard formats such as the Federal Geographic Committee Content Standard for Digital Geospatial Metadata (FGDC CSDGM), the International Standards Organization’s Geographic Information schema (ISO 19115), Ecological Markup Language (EML), and others (Goodchild 2007).

The PDI can be thought of as another set of metadata that is intended to describe information about the preservation and longevity of the CI. PDI describes five categories of information: Provenance, Context, Reference, Fixity and Access Rights (CCSDS 2012). In some cases, these categories may be described within the RI as well—for example, the ISO 19115 metadata standard defines elements for storing metadata within all five of the PDI categories. Regardless of metadata standard, however, some aspects of the PDI cannot be found within the RI because they are not determined until after the creation of the CI. For example, PDI can track information such as when the CI was added to the archive, what users of the archive have access rights to the package, and other facts relevant to the management needs of the archive. The operational details contained in PDI also help to verify the integrity of the archive, for example by storing checksums that alert administrators to changes in the contents of CI, enabling audits of the repository.

A working group co-sponsored by the Online Computer Library Center (OCLC) and the Research Libraries Group (RLG) developed the Preservation Metadata Implementation Strategies (PREMIS) metadata schema specifically for the purpose of implementing preservation metadata in the case of both RI and PDI (PREMIS 2008). The PREMIS metadata standard does not contain elements commonly used to describe geospatial datasets; this is an intentional limitation of scope by the standard's developers due to the existence of geospatial metadata standards listed above and others (PREMIS 2008). Therefore, the RI is better served using a domain-specific standard. However, it may be desirable to describe the PDI using PREMIS metadata, for example in order to standardize the structure of PDI in a repository with heterogeneous metadata using different standards.

The Packaging Information describes the organizational structure of the CI at the computer operating system level—file and directory structure that may be described by ZIP (formalized as ISO/IEC 21320-1:2015) or BagIt (Kunze 2016) archives, or other aggregation schemes. The PD contains information used by data consumers to search for and retrieve the complete Information Package, such as title and abstract fields. The PD can be extracted from the RI and the PDI and inserted into an index to support search and browse functionality.

As the implementation details of the OAIS model are intentionally omitted from the specification, the software design for the repository itself, as well as for related functions, is left as a choice to the architects of such systems. Here, we advocate for the Service-Oriented Architecture (SOA) as an ideal approach to implementation. According to the Reference Model for Service Oriented Architecture developed by the Organization for the Advancement of Structured Information Standards (OASIS), SOA is "a paradigm for organizing and utilizing distributed capabilities that may be under the control of different ownership domains" (OASIS 2006). SOA is a concept from computer sciences that describes building modular, loosely-coupled software systems (Papazoglou and Van Den Heuvel 2006). The modules, or services, that are deployed in such a system may exist in geographically disparate locales; they may be created and maintained by separate institutions or groups, and they may rely on entirely different computing hardware and software. The loose mode of coupling is accomplished through the exposure of an Application Programming Interface (API) that explicitly defines the language and communication protocol through which the service interacts with the outside world. As long as a service properly implements the requirements of the API, it can interoperate with other systems that speak its language. This is opposed to the concept of a "tightly coupled" system, in which components may communicate with each other through channels and protocols that are opaque to outside observers and are generally meant to be invoked only from components within the system, itself.

There are a variety of general motivations for implementing complex software systems using SOA:

- SOA can enhance system reliability: because the system is composed of multiple modules, the failure of any one module does not necessarily mean the failure of the entire repository function, whereas the failure mode of monolithic software systems may bring down the entire suite of functionality (Tsai 2005).
- SOA enables staggered rollout of new features: since service modules (outside a core set of modules) are independent of each other, new features can be implemented as the repository is operating and introduced publicly when they are ready for consumption (Wong-Bushby et al. 2006). In this fashion, an SOA-based OAIS repository can be 'bootstrapped' into a full-featured state over time.
- SOA preserves the functionality of legacy systems: based on the loosely-coupled philosophy of SOA, implementers can design linkages between legacy systems such as institutional/enterprise management software and repositories (Pessoa et al. 2010). As legacy systems transition to more modern versions, linkages can be adjusted to compensate for varying modes of interaction.
- SOA supports interoperability with external systems: similarly to the linkage to legacy systems, loose coupling also supports linkage to systems that exist outside the repository or the institution (Nezhad 2006). Modern systems that are designed for interoperability use standard or well-known APIs that lessen the effort involved in connecting to them from remote systems. Furthermore, repositories designed with interoperability in mind enable catalog and data consumption from external services using standard APIs.

- SOA improves upon the flexibility of monolithic software: one of the challenges of deploying monolithic software solutions is that they are typically designed for a use case that is not exactly reflected within the institution. There may be a component that is missing or unsuited to the environment in which the system is to be deployed. The modular approach of SOA allows implementers to choose individual components from available options, or to implement a particular component themselves (Ren and Lyytinen 2008).
- SOA separates development into manageable tasks: because the modules of an SOA-based repository take advantage of loose coupling and APIs to interact with each other, maintenance, bug fixes, and development work done on one component do not immediately require making changes to the internal code of another. If new functionality is required of the repository, the functionality can be implemented one component at a time, reducing the complexity of development tasks. When APIs are updated with new functionality, older functionality can be maintained by continuing to support older versions of the API (Josuttis 2007). This can help to maintain links to legacy and external systems.
- SOA allows distribution of repository functions across geography and institutions: as interdisciplinary research and large-scale collaboration increase in popularity, it is important that data management systems are able to federate functionality and content with each other (Yarmey 2014). Even standards-based repositories do not always follow the same standards, especially across international borders. The SOA approach to interoperating with external systems can be crucial for communication across institutions.
- SOA allows the compartmentalization of user access rights and security (Channabasavaiah et al. 2003). Since each service operates using its own security model and user authentication requirements, privileged access can be reserved for users and modules that definitely require heightened levels of access.
- SOA helps to avoid problems associated with vendor lock-in (Brown 1998). With monolithic software, administrators face deadlines such as end-of-life dates, at which the entire software package must be upgraded to a newer version, regardless of whether the newer version represents an improvement over the old one for users.

A theme that emerges among the strengths of SOA is ease of adapting to change. In order to provide value, the continuing development of science data repositories must be driven by the dynamic needs of the research communities that feed them.  The data deluge involves research products that are growing in size and complexity faster than existing systems can accommodate (Hey and Trefethen 2003).  Repositories must be prepared to adapt to support data of various scales, from small legacy text-based data to newer terabyte- or higher-scale collections.  New science and technologies often involve data stored in novel file or database formats (Ahrens 2011).  As these novel formats proliferate, they enable an increasingly heterogeneous list of new features and capabilities, pushing repositories to expose new and updated services.  As repositories are driven to federating and other methods of interoperability, they must adapt to the choices and limitations of technologies implemented by potential partners.  These and other adaptations are strongly supported by the SOA approach.

There are also limitations to the SOA approach to developing repositories.  The need to adapt SOA-based repositories to accommodate new conditions represents engineering challenges for software developers (Palma 2013).  Keeping the various services of the system functioning and interoperating smoothly can be another challenge.  Relying on monolithic software allows repository administrators to focus on the business of managing and curating data, rather than overseeing the continued development and maintenance of software services.

There are a variety of monolithic, off-the-shelf software choices for repositories.  According to the Registry of Research Data Repositories, which surveys research data repositories worldwide, out of 1,763 repositories, the top three data management systems are DSpace (42 instances), DataVerse (36 instances), and CKAN (28 instances) (Re3Data 2016).  These numbers likely underestimate the number of repositories using these software packages—the vast majority (1,266 instances) are listed as either "other" or "unknown".  Amorim (2016) presents a more complete list of repositories and performs some evaluation of their relative merits.  Some, like DSpace, specifically aim to implement the OAIS model, but most do not.

Adherence to the OAIS model for repositories comes with several advantages.  First, OAIS-based repositories take advantage of the deep thought and planning by a large body of researchers that has gone in to building the model.  Second, the CCSDS is developing a recommended practice for the Audit and Certification of Trustworthy Digital Repositories "to create an overall climate of trust about the prospects of preserving digital information" (CCSDS 2011).  Furthermore, the application of a standard model may serve to improve interoperability between repositories due to the use of common paradigms in design and implementation.

The OAIS model explicitly “does not specify a design or an implementation” (CCSDS 2012). In part due to this, and also due in part to the uncertain speed, reliability, and persistence of Internet connections during the inception of the OAIS model, one area that is not well-developed is the connection of data repositories to network- or cloud-based services and resources. We use the Service Oriented Architecture (SOA) design paradigm to describe a set of extensions to the OAIS Reference Model that enable a repository to take advantage of recent opportunities for interoperability. We describe a purpose and justification for each extension, where and how each extension connects to the model, an example of a specific implementation that meets the purpose, and a suitable API definition to support the functional purpose.

## Methods

### *Data Unique Identifiers*

In order for data consumers to make use of data in repositories, the data must have a persistent point of access and must be verifiably the data that the consumer is interested in. Unique identifiers for data can provide access keys that decouple location information from identifiers so that when data are moved, identifiers remain consistent while location information is updated in linked databases. By maintaining a consistent identifier for data, citations in publications and other documents resist becoming stale, so consumers can maintain access to data and be sure they are accessing the data they are expecting.

In 2011, Duerr et al. published an assessment of nine different data identification schemes: ARKs, DOIs, XRIs, Handles, LSIDs, OIDs, PURLs, URIs/URNs/URLs, and UUIDs. Of these, the Digital Object Identifier (DOI) stands out as a strong candidate for application in data repositories given its widespread adoption by publishers, its acceptance as an ISO standard (ISO 26324), and its interoperability with other common location and identification schemes such as Uniform Resource Identifiers (URIs). The DOI is a product of the International DOI Foundation “designed as a generic framework applicable to any digital object, providing a structured, extensible means of identification, description and resolution (International DOI Foundation 2012).”

The DOI works via a central registry that associates unique identifiers with the locations of data products. The recommended practice for assigning the endpoint of a DOI is not to link directly to data products, but to web pages that display descriptive information about the data products, often to include download links or instructions for obtaining the data if not available for download (International DOI Foundation 2012). This descriptive information can be derived directly from a repository's representation information, providing the consumer with some certainty that they have found what they are looking for.

The infrastructure requirements for accommodating the DOI are modest. It requires a method of storing the DOI value such that it is associated with the data object that it identifies, a method of discovering DOI values that are stored within the repository, and a method of resolving client requests for DOIs.

The ideal storage location of the DOI is within the metadata associated with a data object, as this identification information is solidly within the purview of the purpose of Representation Information. However, not all metadata standards allow for the storage of a DOI in an unambiguous way. The FGDC CSDGM, for example, defines no specific location for the storage of a DOI. Although one could be stored in a variety of locations within a metadata satisfying the standard, the weak semantic cues given by more general-purpose fields makes it difficult for an automated process to identify unambiguously that a DOI that it finds within them is the correct identifier for the dataset. To account for cases in which the representation information standard does not allow for unambiguous storage of the DOI, the PDI can also be used to store the DOI using a standard such as PREMIS or an ad-hoc approach.

The discovery method for the DOI can be as simple as for any other field within the metadata: index the DOI field and present it through the normal search interface.

The data unique identifier module integrates with the OAIS model in three places: at the ingestion phase, where a user can input or assign the DOI information relevant to the record being inserted; in the storage system, where the Packaging Information associates the DOI with the repository record; and at the access phase, where users can query the repository based upon the DOI.

Some issuers of DOIs, such as the California Digital Library's (CDL) EZID service (http://ezid.cdlib.org/), expose an API that allows clients to request a DOI for a data object. At the ingestion phase then, the repository accesses the remote API to issue a new DOI for the ingested data object and inserts the DOI into the Representation or Preservation Description Information for storage.

Resolving client requests for access to data identified by DOIs involves accepting a query for a DOI; looking up the DOI in the repository; and either retrieving and rendering a search result, or indicating the failure of the repository to resolve the DOI. Following the best practice for DOIs resolving to descriptive landing pages, the repository may generate a page upon request, based upon information from the RI and PDI. The dynamic generation of a landing page based upon the object's metadata helps ensure that landing pages always include the most up-to-date, authoritative description available for the Information Package.

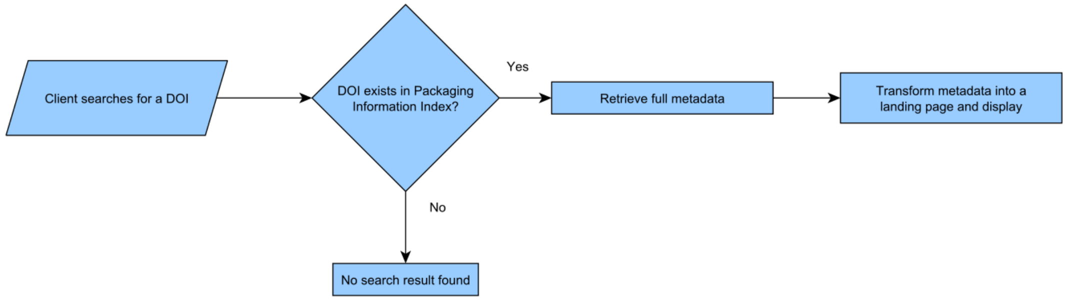
Client searches for a DOI
DOI exists in Packaging Information Index?
Yes
Retrieve full metadata
Transform metadata into a landing page and display
No
No search result found

Figure 2.3 Example of a consumer interacting with a repository by requesting details of a DOI

### *Researcher Unique Identifiers*

One common difficulty in the academic publishing arena is the potential for ambiguity of authors' names. A researcher may, over the course of a career, publish under more than one name, making it difficult to assemble an exhaustive list of their publications. Multiple researchers may share the same name (or initials), making it difficult to separate their individual bodies of work. The combination of a researcher name and institution can help, but is still problematic when researchers change employers, have multiple appointments, or have common names and are associated with large institutions (Han et al. 2004).

One approach to disambiguating author names is to associate unique identifiers with authors. This approach requires a certain amount of cooperation between authors, who must agree to participate in a registry and keep their record up-to-date; publishers, who must agree to include the unique identifiers with publications; and a central authority that maintains the registry of mappings between authors and their unique identifiers. One such organization that has seen widespread adoption is Open Researcher and Contributor ID (ORCID), which operates a web-based registry (Haak et al. 2012).

The author identifier service has three useful points of interaction with the OAIS research data repository. The first is when the data producer initiates the ingestion process. The producer will be requested to create a certain amount of metadata to describe the data that they are submitting to the archive. As part of that metadata collection effort, the producer should be given the opportunity to provide their unique identifier.

As with the DOI data identifiers, most representation information standards have no explicit way to store ORCID or other systematized researcher IDs. The ORCIDs may be stored in a variety of ways within metadata, but are difficult to store in a semantically unambiguous way.

The ORCID API allows systems to query the ORCID database to retrieve public data about authors who are indexed in the system. A researcher unique identifier module, then, can interface with the OAIS repository in three ways.

The second point at which the author identifier can interact with the repository is during the ingestion process, when the ORCID is collected from the producer. The repository can use the ORCID API to query for and populate fields related to producer identification using the data that are publicly available from the ORCID database. This step can save time and effort for the producer by obviating the need to manually enter simple identification information.

At the storage phase, the repository then stores the ORCID in a designated field within the Packaging Information associated with the data object. As a part of periodic metadata maintenance, it is then possible to compare the stored ORCID for a data object with the producer information stored within the representation information and check for mismatches. It is not clear in these audits whether the metadata has fallen out of sync with the reality that is represented in ORCID or the other way around (alternatively, both the representation information and ORCID database may have become obsolete), but it is at least possible to use the audit to flag the record for a human to review and try to find a resolution.

The third point of interaction between the author identifier and the repository is at the data consumer interface, when a potential consumer wishes to search for data produced by a particular researcher. If the consumer is able to search using the producer’s unique identifier as a key, the results that they retrieve should be unambiguous. As with the DOI, the discovery method for the ORCID can be as simple as for any other field within the metadata: index the ORCID field and present it through the normal search interface.

### *Federated User Credential and Identity Management*

With today’s focus on interdisciplinary research projects that can span multiple institutions, researchers can face challenges in dealing with disparate information technology systems. One such challenge is user credential management—while each participant in a research project has a set of computer credentials issued by their institution, these credentials are rarely interoperable. That is, computer users at one institution cannot use their login credentials with systems at another institution.

These incompatible credentials lead to problems with the research data repository. The repository may support a standalone credential system, but how can repository operators know whether users from other institutions are allowed certain types of access? Even if users have authenticated with their home institutions, how can these credentials be trusted? On-line identity theft is a growing problem in the business sector, and can easily transition to the research world. Some research data may be protected by statutes such as FERPA or HIPAA, some may be protected by agreement with an institutional review board, and some may be sensitive due to their unique nature; it is therefore important to maintain a system of credential management for repository users in order to control access and management of data assets.

A potential solution to this issue is federated credential management, a system in which institutions join together to vouch for the validity of their users' login credentials (Bhatti et al. 2007). There exist a variety of organizations providing federated credential services, many focused on particular geographic areas or activity domains. A popular provider among academic institutions in the United States is InCommon (Barnett et al. 2011). These organizations allow credential providers to issue usernames and passwords to their users and to share their authentication process with external systems without transmitting or revealing the actual credentials. In this way, individual institutions can continue to manage the basic details of user credentials such as login names and passwords while enforcing their own local policies. Federated credentials can interact productively with researcher unique identifiers as well: if credential stores contain ORCID information, and expose that information to systems consuming their authentication services, then federations can share not only credentials, but also identities.

From the repository perspective, managers can grant access rights to users based upon information gleaned from the federated credential service. Based upon common identifiers such as ORCID, repository managers can arrange permissions to allow individual users or groups of users to create, modify, or view data packages stored in the repository.

The federated identity system can connect with the OAIS model at any point of connection into the archive from outside: producer, consumer, or manager. The mode of connection is through the API exposed by the identity management system. This API is responsible for accepting authentication credentials and returning some base level of information about the user that has successfully logged in: at the minimum, a user ID that is compatible with the local repository system. Ideally, more information would be shared: user data such as ORCID and other descriptive data that help the repository to categorize the external user. Once a user has authenticated, access rights can be managed just as with any traditional, locally existing user. In this way, multiple repositories can share user identities without the need of sharing user credentials, and can grant privileges within the repository to users of other repositories to support collaboration across institutions. When identity information is included in addition to credentials, external users gain the benefits provided by the repository's researcher unique identifier module.

### *Harvesting, Federated Catalogs, and Search*

Given the proliferation of research data repositories—Marcial and Hemminger (2010) identified thousands of science data repositories in a survey in 2010—potential data consumers may not be aware of repositories that could hold information that would further their research goals. Rather than searching many repositories individually, it can be helpful to the user to be able to search many repositories simultaneously.

Two related approaches to expanding search capabilities to the content of multiple repositories are harvesting and federated search. Harvesting is a process of collecting remote metadata records into the local repository, ingesting them automatically for search. Federated search involves applying search terms not only to the local repository, but also to remote repositories to find results. For both of these approaches, a repository must provide a means of both supplying and consuming these services.

One popular way of arranging harvesting services is through the standard protocol, Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). In order to harvest data from a repository that implements OAI-PMH, a harvester makes a "ListRecords" request and the remote repository responds with an OAI-PMH envelope that contains a series of records that list an identifier, a datestamp, and a metadata record conforming to a format specified in the request.

The first two of these items should be readily available from the packaging and representation information. The metadata may be more difficult to come by if the requested format is not the native format of the representation information. The bulk of the response implementation, then, is implementing some translation service that can produce at least a minimal metadata record based upon information found in the representation information. Since this metadata will only be used for data discovery purposes, only a small number of fields must be populated; the difficulty may arise from diversity of source locations for the content of these fields based upon the variety of representation information standards that are stored within the repository.

Federated Search is one approach for addressing this need. Federated Search provides the user with one search interface that connects with many back-end repositories and provides results in aggregate form (Shokouhi 2011). Federated Search is most easily accomplished when repositories offer a common search API, obviating the need for custom computer code to connect to different repositories. One common federated search protocol is the Open Geospatial Consortium's Catalog Service for the Web (OGC CSW) (Liakos et al. 2015).

Within the OAIS model, the standard search API would be implemented at the Access block that interfaces with the data consumer. Multiple access methods may be implemented, and there are several data repository systems that support multiple standards and ad-hoc methods of access.

At the heart of supporting federated search is exposing some amount of the packaging and representation information to other systems using a well-known API. Regardless of the metadata standards and implementations used within the repository, if the necessary details of data objects can be organized into valid API responses, then the data can be made searchable through federation.

### *Data Object Replication*

Another set of challenges for potential data consumers can be dealing with slow transfer speeds resulting from long geographic (or network topological) distance to the repository, and data that are inaccessible due to repository or network down time. If users are unable to achieve reliable access to archived data, they are unlikely to rely on such data for their own research purposes.

One method for mitigating the risks of low-availability data is to replicate the data in multiple disparate geographic areas, decentralizing risk across networks and nodes. This can be done at several conceptual levels within the repository architecture. For example, the "Archival Storage" that the data consumer accesses may not be a single storage system, but a distributed file system such as can be accessed through the Amazon Simple Storage Service (S3) (https://aws.amazon.com/s3/). In a replication system implemented at that level, the repository itself need not be aware of the particulars of the geographic locations of files; the file system is abstracted sufficiently from the repository that at any geographic (or network-topological) location, a data consumer who accesses a data object is automatically given access to the nearest copy.

One approach to this replication is the NSF-funded DataONE project, which uses an OAIS-like implementation to ingest data into member nodes and then distribute replicas of data objects to several other member nodes in other locations around the network (Reichman et al. 2011).

In this mode of replication, the repository may need to be more involved. Data object replication can be thought of as a scenario in which an agent consumes data objects from one repository and produces those same objects for ingestion into a second repository. In order for data replication to occur in an automated and predictable way, a common data access API can be implemented at the Access block that interfaces with the data consumer. A data ingestion API can be implemented at the Ingest block of the model that interfaces with the data producer. In this case, a software agent interfaces with these APIs to connect two repositories. Such an agent may be operated by one or the other (or both) of the endpoints of the replication transaction and may require some supporting Packaging Information to be associated with the data objects, for example to indicate that a particular data object is an ideal candidate for replication.

A further benefit of replication is that it provides some redundancy of data object storage that can make data more robust against catastrophic events. Should one repository be struck by an irrecoverable data loss scenario, data that have been replicated to other sites should still be available. Though replication is not equivalent to, and should not be used in lieu of, a traditional backup system, it may serve a similar purpose.

### *Version Control*

As time passes, information contained in metadata tends to fall out of date, particularly in the case of information about people and institutions associated with data—names, phone numbers, addresses, the organization of institutions. These details tend to change over time. Much more rarely, changes will need to be made to the sections of metadata referring to the data, themselves. In either case, as changes are made to metadata records, it can be difficult to compare two metadata records and determine whether or not they describe the same dataset and are, in fact, two different versions of the same metadata record.

The issue of data provenance is important when considering using research data secondarily. It is critical that a researcher knows if changes have been made to a data object since its creator first published it into an archive, both to determine the data's suitability for use and to be able to accurately represent the full extent of data processing methods that have been applied.

Version control systems (VCS) offer the capacity to look back at previous versions of files that are stored within them and see in precise detail how those files have changed over time (Sen 2004). There are several popular version control systems today; foremost among them are Git (https://git-scm.com/) and Subversion (https://subversion.apache.org/). Version control can be deployed within an OAIS compliant repository's Archival Storage system.

For example, a VCS such as Subversion can operate in a way that is mostly transparent to the repository except when its special functions are needed. The repository continues to keep metadata in its usual storage system, registering each record with the VCS. As metadata records are updated by producers, the VCS keeps a history of each record, tracking changes to the metadata. Consumer users of the repository are presented with the latest version of a metadata record by default, but on request, the VCS can provide a detailed revision history. For consumers who are interested in previous versions of metadata records, there are many existing tools that allow powerful browse and search capabilities, such as the open-source Windows application, TortoiseSVN (https://tortoisesvn.net/). Using such tools, a data consumer can check previously downloaded representation information against old versions stored within the VCS to verify that they are using an older version of the same data object.

For repository administrators, the version control system provides an audit trail that allows them to identify who has made changes to a file, at what time, and of what substance. This information can be used in the development of detailed provenance records for data and metadata. Reporting on update activity can also give administrators insight into how data producers are interacting with the repository, which metadata records undergo frequent update, and why—potentially helping to inform the kinds of training and assistance offered to producers. The VCS also grants administrators the capability of inspecting and reverting changes that have been applied erroneously as metadata records are maintained. Like data replication, VCS can offer a kind of backup capability for the repository, allowing damage to be undone

### *Taxonomies and Controlled Vocabularies*

Taxonomy services provide access to controlled vocabularies for use by organizations and disciplines to classify things (Cohen 2007). In data management, the controlled vocabulary can be used to provide a consistent set of descriptive terms used to describe a dataset. Consistency enhances the ability of search clients to be able to locate records described by a particular term. For example, when using keywords to describe geographic data collected within the United States of America, it is useful to have a common term such as “USA” rather than a proliferation of variations such as “U.S.A.”, “US”, “U.S.”, “United States”, “America”, etc.

A wide variety of taxonomy services exist, particularly services suited to certain research domains. The ISO 19115 Topic Categories is a simple example of a taxonomy intended to describe a general theme of geospatial data. It contains only 19 terms: farming, biota, boundaries, climatologyMeteorologyAtmosphere, economy, elevation, environment, geoscientificInformation, health, imageryBaseMapsEarthCover, intelligenceMilitary, inlandWaters, location, oceans, planningCadastre, society, structure, transportation, and utilitiesCommunication (ISO 2007). The generality and limited number of terms of this taxonomy limit the ability to express complex information about a dataset, but do provide a standard set of terms that may be used in data discovery.

The USGS Geographic Names Information System (GNIS) (http://geonames.usgs.gov/) is a much more elaborate taxonomy that records the variety of official names for geographic features across the United States. Containing more than two million entries, this taxonomy can be used to specifically identify a geographic location, but can prove daunting as a search tool due to its size.

Taxonomy services are useful at two stages in the data ingestion process. First, the data producer can take advantage of the service while producing the metadata record. This application is beyond the scope of the data repository itself, but metadata creation/editing utilities may be designed to interface directly with the repository, so it can be of benefit to coordinate between any utilities created and any repositories used to ensure that they use common taxonomy services.

The second application of the taxonomy service occurs in the Quality Assurance block of the Ingest system. If the repository mandates the use of certain taxonomies where applicable in metadata, then the QA process can use the taxonomy service to verify the content of the relevant metadata elements, rejecting non-complying metadata for further review by producers.

### *Live Data Exposure*

As the resolution of measurements across many dimensions increases with access to advanced instruments and massive storage systems, data consumers may prefer not to copy entire data sets for local use, instead opting to extract useful subsets or aggregations of data, or simply to connect to services that expose data and perform analyses remotely. When dealing with very large data collections, it makes sense to transfer only those parts of the data that are involved in analysis in order to conserve transfer time, local storage, and computational resources used in analysis.

To that end, data services such as the OGC Web Feature Service (http://www.opengeospatial.org/standards/wfs), the Unidata Thematic Realtime Environmental Distributed Data Services (THREDDS) Data Server (http://www.unidata.ucar.edu/software/thredds/current/tds/), the Consortium of Universities for the Advancement of Hydrologic Science Hydrologic Information Service (CUAHSI HIS) (http://his.cuahsi.org/), and others have arisen. These services provide a consumer-facing API that accesses the Archival Storage block to manipulate and expose data in formats that are friendly to the consumer's data client software, where they may then be analyzed and visualized as if they were local resources.

For many of these services to function, most of the repository system need not be involved directly. As an example, the THREDDS service can be set up as a new Access component to the repository; THREDDS becomes a new "Live Access" component of the repository, another way for the consumer to access the data.

## Conclusion

The OAIS reference model is intentionally devoid of implementation detail, but our current climate of cloud- and network-based services lends itself to low-level interaction with some internal parts of an OAIS repository. We have described unique identifiers for data that help to provide long-term access and assure the identity of the data object; unique identifiers for researchers that disambiguate data producers and can help to identify a researcher's body of work; federated user identity management that provides a single set of credentials that enable access controls; federated catalogs and search that help make data objects accessible through more interfaces and to more potential consumers; data replication that can provide redundancy protection against certain kinds of data disasters and enables fast access to data objects by consumers; version control that provides audit histories of metadata that allow for the comparison of metadata records; taxonomy services that help to control search vocabularies to help consumers search for data; and live data services that can obviate the need for data consumers to download large data objects in situations where they may only need small parts. Together, these services serve as a set of implementation details for an OAIS repository that are relevant to our modern level of connectivity and collaborative research.

## References

Ahrens J, Hendrickson B, Long G et al (2011) Data-intensive science in the US DOE: case studies and future challenges. Computing in Science & Engineering 13(6):14-24

Amorim RC, Castro J A, da Silva JR, Ribeiro, C (2016) A comparison of research data management platforms: architecture, flexible metadata and interoperability. Universal Access in the Information Society, 1-12. doi:10.1007/s10209-016-0475-y

Barnett W, Stewart CA, Walsh A, Welch V (2011) A roadmap for using NSF cyberinfrastructure with InCommon. http://hdl.handle.net/2022/13024 and http://www.incommon.org/nsfroadmap.html. Accessed 13 December 2016. doi:2022/13024

Bhatti R, Bertino E, Ghafoor A (2007) Federated identity and privilege management. Commun ACM 50(2): 81–88. doi:10.1145/1216016.1216025

Brown WJ, Malveau RC, McCormick HW III et al (1998) AntiPatterns: refactoring software, architectures, and projects in crisis. John Wiley & Sons, New York

CCSDS: Consultative Committee for Space Data Systems (2011) Audit and certification of trustworthy repositories. https://public.ccsds.org/pubs/652x0m1.pdf. Accessed 13 December 2016

CCSDS: Consultative Committee for Space Data Systems (2012) reference model for an Open Archival Information System (OAIS). https://public.ccsds.org/pubs/650x0m2.pdf. Accessed 13 December 2016

Channabasavaiah K, Holley K, Tuggle E (2003) Migrating to a service-oriented architecture. IBM DeveloperWorks 16

Clarke L, Zheng-Bradley X, Smith R et al (2012) The 1000 genomes project: data management and community access. Nat Methods 9.5:459-462. doi:10.1038/nmeth.1974

Duerr RE, Downs RR, Tilmes C et al (2011) On the utility of identification schemes for digital earth science data: an assessment and recommendations. Earth Science Informatics 4:139. doi:10.1007/s12145-011-0083-6

Cohen S (2007) Ontology and taxonomy of services in a service-oriented architecture. Microsoft Architecture Journal 11:30–35

Goodchild MF (2007) Beyond metadata: Towards user-centric description of data quality. Keynote paper, Proceedings, 5th Int. Symposium Spatial Data Quality, ITC, Netherlands, 13–15 June

Gunelius S (2014) The Data Explosion in 2014 Minute by Minute. http://aci.info/2014/07/12/the-data-explosion-in-2014-minute-by-minute-infographic. Accessed 13 December 2016

Haak LL, Fenner M, Paglione L et al (2012) ORCID: a system to uniquely identify researchers. Learned Publishing 25(4):259–64. doi:10.1087/20120404

Haeberli C, dos Anjos A, Becket HP et al (2004) ATLAS TDAQ data collection software. IEEE Transactions on Nuclear Science 51(3):585–590

Haendel MA, Vasilevsky NA, Wirz JA (2012) Dealing with data: a case study on information and data management literacy." PLoS Biol 10.5:e1001339. doi:10.1371/journal.pbio.1001339

Han H, Giles L, Zha H et al (2004) Two supervised learning approaches for name disambiguation in author citations. Proceedings of the 2004 joint ACM/IEEE conference on Digital Libraries, pp 296–305

Hey AJG, Trefethen AE (2003) The data deluge: an e-science perspective. In: Berman F, Fix GC, Hey AJG (ed) Grid Computing: Making the Global Infrastructure a Reality. Wiley, New York, pp 809–824

Hilbert M, López P (2011) The world's technological capacity to store, communicate, and compute information. Science 332.6025:60–65. doi:10.1126/science.1200970

International DOI Foundation (2012) DOI Handbook. http://www.doi.org/hb.html. Accessed 13 December 2016

ISO: International Organization for Standardization (2007) ISO/TS 19139:2007: Geographic information–Metadata–XML schema implementation. http://www.iso.org/iso/catalogue_detail.htm?csnumber=32557. Accessed 13 December 2016

Josuttis NM (2007) Versioning. In: St. Laurent S (ed) SOA in practice: the art of distributed system design. O'Reilly Media, Sebastopol CA, pp 145–157

Kunze J, Boyko A, Littman J et al (2011) The bagit file packaging format (v0. 97). https://tools.ietf.org/html/draft-kunze-bagit-08. Accessed 13 December 2016

Lavoie BF (2014) The Open Archival Information System (OAIS) reference model: introductory guide (2nd Edition). www.dpconline.org/component/docman/doc_download/1359-dpctw14-02. Accessed 13 December 2016

Liakos P, Koltsida P, Kakaletris G et al (2015) A distributed infrastructure for Earth-science big data retrieval. International Journal of Cooperative Information Systems 24(02):1550002. doi:10.1142/S0218843015500021

Marcial LH, Hemminger BM (2010) Scientific data repositories on the web: An initial survey. J Am Soc Inf Sci Technol 61(10):2029–2048. doi:10.1002/asi.21339

Nezhad HRM, Benatallah B, Casati F, Toumani F (2006) Web services interoperability specifications. Computer, 39(5):24–32

OASIS: Organization for the Advancement of Structured Information Standards (2006) Reference model for service oriented architecture version 1.0. http://docs.oasis-open.org/soa-rm/soa-ra/v1.0/cs01/soa-ra-v1.0-cs01.html. Accessed 13 December 2016

Palma F, Nayrolles M, Moha N et al (2013) SOA antipatterns: an approach for their specification and detection. International Journal of Cooperative Information Systems 22(04):1341004

Papazoglou MP, Van Den Heuvel WJ (2006) Service-oriented design and development methodology. International Journal of Web Engineering and Technology 2(4):412–442

Pessoa RM, Silva E, van Sinderen M et al (2008) Enterprise interoperability with SOA: a survey of service composition approaches. 2008 12th Enterprise Distributed Object Computing Conference Workshops, pp 238–251

PREMIS: PREMIS Editorial Committee (2008) PREMIS data dictionary for preservation metadata version 2.0. http://www.loc.gov/standards/premis/v2/premis-2-0.pdf. Accessed 13 December 2016

Reichman OJ, Jones MB, Schildhauer MP (2011) Challenges and opportunities of open data in ecology. Science 331(6018):703–705. doi:10.1126/science.1197962

Re3data: Registry of Research Data Repositories (2016). http://www.re3data.org/. Accessed 11 August 2016

Ren M, Lyytinen KJ (2008) Building enterprise architecture agility and sustenance with SOA. Communications of the Association for Information Systems 22(1):4

Sen A (2004) Metadata management: past, present and future. Decision Support Systems 37(1):151–73. doi:10.1016/S0167-9236(02)00208-7

Shokouhi M (2011) Federated search. Foundations and Trends in Information Retrieval 5(1):1–102. doi:10.1561/1500000010

Tenopir C, Allard S, Douglass K et al (2011) Data sharing by scientists: practices and perceptions. PloS One 6(6):e21101. doi:10.1371/journal.pone.0021101

Tsai, WT (2005) Service-oriented system engineering: a new paradigm. Proceedings of the 2005 IEEE International Workshop on Service-Oriented System Engineering, pp 3–6

Wong-Bushby I, Egan R, Isaacson C (2006) A case study in SOA and re-architecture at company ABC. Proceedings of the 39th Annual Hawaii International Conference on System Sciences

Yarmey L, Khalsa SL (2014) Building on the international polar year: discovering interdisciplinary data through federated search. Data Science Journal 13(0):PDA79-PDA82

Zikopoulos P, Eaton C, deRoos D et al (2011) Understanding big data: analytics for enterprise class hadoop and streaming data. McGraw-Hill, New York

# Chapter 3: Building an Open Science Framework to Model Soil Organic Carbon

**Flathers, E., & Gessler, P.E. 2018. Building an Open Science Framework to Model Soil Organic Carbon. Journal of Environmental Quality. Special Issue on: Predicting Soil Carbon in Agroecosystems Under Climate Change. doi: 10.2134/jeq2017.08.0318**

## Introduction

As funding bodies for research (e.g. USDA, National Science Foundation, National Institutes of Health, among others) embrace free and open publication of research data, many science disciplines are developing a new culture of data sharing. For example, a recent study using magnetic resonance imaging data has uncovered a flaw in a common analytical technique. The authors note: “through the introduction of international data-sharing initiatives in the neuroimaging field, it has become possible to evaluate the statistical methods using real data” (Eklund et al., 2016). Data sharing has enabled scientists to check methods and improve methods in ways that haven’t previously been possible allowing rigorous science to advance more quickly.

In addition to data sharing, researchers are developing software systems to help enable a more complete sharing culture. For example, the Center for Open Science developed the Open Science Framework (https://osf.io) to organize components of research projects and enable collaboration and sharing of project materials including data, code, and text both during and after projects. The adoption of open science practices, including sharing research data and software, has the potential to improve the progress of science in the same way that publication of methods and results in journal articles, as scientists learn and take inspiration from the works of their peers. But open science practices can also improve science in other ways. Gezelter (2015) argues that “as numerical experiments become more complex and the data sets become larger, calculations that are reproducible in principle are no longer reproducible in practice without access to the code, data, and the meta-data that describes how the data is organized”.

A 2016 Nature survey showed that 52% of researcher respondents (no information was given about their fields of study) agreed there is a “significant crisis of reproducibility” across the sciences (Baker, 2016). The Reproducibility Crisis is caused by factors common to all science: “Problematic practices include selective reporting, selective analysis, and insufficient specification of the conditions necessary or sufficient to obtain the results" (Open Science Collaboration, 2015). Openly sharing science data, code, and products provides an avenue for reproducing results in any discipline (McNutt, 2012).

One obstacle to the embracing of open data sharing is the danger of being scooped by other researchers who use shared data to publish papers. This type of scooping has allegedly occurred in genomics, when MIT researchers published a paper using data made publicly available by the Woods Hole Marine Biological Laboratory in a way contrary to the restrictions imposed upon the data by the authors (Marshall, 2002). One approach to mitigating this danger is a publication embargo, which grants original researchers time to publish using their data before they become public (Cragin et al., 2010). However, the length of the embargo should not extend beyond the useful life of data and code. Ideally the embargo provides researchers a head-start on publication while also allowing access to others while products are still relevant.

Reproducibility is required for rigorous science and demonstrates the fundamental stability of the methods applied in a study. Without the ability to reproduce an experiment, scientists have no way to judge the veracity of the results. Asendorpf et al. (2013) define reproducibility as providing a set of outputs that researchers must produce to enable reproduction of their studies: raw data, metadata, and the actual code used to perform the analyses.

Based upon this definition of reproducibility, this paper documents a dataset and associated analytical products that meet Asendorpf's criteria. The USDA-funded Regional Approaches to Climate Change for Pacific Northwest Agriculture (REACCH-PNA) is a project focused on the potential impacts of climate change on cereal grain production in the northwestern United States. One product for the study area is a derived map of SOC as a base from which C dynamics can be mapped and monitored. Soil organic C is primarily associated with soil organic matter and relates to many soil properties that influence resiliency and soil health. It is also critical for understanding soil-atmospheric C flux, which is a significant part of the overall C budget of the Earth (Raich and Schlesinger, 1992). Though there are ongoing efforts to produce global maps of SOC (FAO, 2017), the scope of this model is limited to a smaller geographic area.

The soil C map is produced by applying a scorpan technique to create a random forest statistical model to predict SOC content for a spatial grid of 30 m resolution. The scorpan model is a more recent implementation of Jenny's (1941) quantitative pedology work into a framework for predicting soil types and properties (Florinsky, 2012). Calibration and evaluation of the model is performed using point-based SOC observations. The explanatory variables are gridded geospatial data describing soil, climate, organisms, topography (relief), parent material, age, and spatial position (McBratney et al., 2003). Because soil respiration creates a flux of soil C that is partly dependent upon an erosion/deposition cycle (Doetterl et al., 2016), topography-derived hydrological and geomorphological layers are also included in the model (Gessler et al., 2000). All inputs to the model are collected from data that various agencies and researchers have made freely available on-line.

"Big data" is an overloaded term in research—it is used to describe data that are large in volume, variety, velocity, value, or complexity (Kaisler et al., 2013). Though the explanatory variables involved in the scorpan model are large (approximately 180 GB), the volume of data is not the greatest challenge in collating the inputs. The more significant big data challenge is the variety of data: a collection of geospatial data produced by a diversity of organizations at different times and different spatial and temporal scales, using varied units of measurement—to name just a few of the differences. The Extract, Transform, Load (ETL) process is designed to ease the integration of the data (Vassiliadis, 2009). Despite this, some artifacts of the disparate origins of the data remain. The "ecological fallacy" describes a misinterpretation of statistical data in which a characteristic of aggregate data is simply applied to groups within the aggregate (Selvin, 1958; Piantadosi, 1988). Because some of the gridded input data for the model are collected at larger spatial scales (4 km cells vs 30 m cells), we commit the ecological fallacy when we assign the attributes of a 4 km cell to the 30 m cells contained within. Despite this, we proceed with the analysis as a real-world compromise; it is often the case that data do not fit elegantly together for analysis, and it is important to be explicit about potential sources of error.

The model output product layers can be used as inputs to other spatial analysis projects (Moore et al., 1993; Gessler et al., 1995; 2000). Instruments located around the REACCH-PNA study area are monitoring carbon dioxide (CO2)  flux, and combining stored soil C data with CO2 flux data for a better understanding of how soil-atmospheric  CO2 flux relates to stored soil C, and how fluxes change over time. Additionally, other members of the REACCH-PNA team are working on projects such as CropSyst, a crop simulation model that outputs, among other attributes, soil organic matter (Stockle et al., 2003). Soil C maps have the potential to provide inputs or serve as a basis for comparing inputs and outputs of CropSyst and other models. These maps also provide for combination with climate change scenarios and known agro-ecosystem domains that suggest shifting of cropping systems as a result of climate change. This effort helps develop the building blocks for such analyses across the region.

The aim of this paper is to create and demonstrate a repeatable, re-usable framework for applying a scorpan model for mapping SOC over the REACCH study area. This is an initial step toward developing accurate spatially explicit soil C maps to support analyses understanding that the initial map is likely inaccurate. Explicit publication of data and methods provides a framework to refine the modeling and improve the outputs. The focus is to demonstrate the concepts of open science and a re-usable and modifiable framework that can be improved upon or applied in other spatial and temporal contexts and scales. All modeling components including input data, metadata, computer code, and output products are made freely available under an explicit open source license. In this way, Asendorpf's criteria are explicitly met; the methods and code released are available for re-use; and research products are plainly open to critical review and improvement.

## Data development process

### *Software and algorithms*

The methods listed in this section are all implemented in the statistical programming language R (https://www.r-project.org/) and the more general-purpose programming language Python (https://www.python.org/) combined with the Esri ArcGIS Python Application Program Interface (API) (https://developers.arcgis.com/python/) and Spatial Analyst (http://www.esri.com/software/arcgis/extensions/spatialanalyst). The ArcGIS code invoked here is used only during the ETL process for tasks such as clipping and projecting raster data, while the analytical work is done using R. Ideally, all of the software code invoked by this project would be open-source and available for audit, but the proprietary Esri ArcGIS was chosen for the convenience of its availability and implementation. Other GIS software, such as GRASS GIS (https://grass.osgeo.org/), would be an ideal alternative, given its open-source implementation. The Python programming language has seen widespread adoption in the environmental sciences, and has been developed under an open-source license (Lin, 2012). The Python code used to implement processing and derivation methods for each input layer is made available on GitHub (https://github.com/), a popular web site used for sharing open-source software code. The Python code is also commented with citations to papers that describe the processing methods.

### *Collected Data Products*

The input products downloaded for use in the model are listed below. The general workflow for each data input is to

Download data from provider

Pre-process the data

Project, if necessary, into Albers Equal Area projection

Clip data to remove any cells outside the REACCH bounding box

Write the resulting raster to disk

Two of the input data products involve processing beyond this general workflow. Using the digital elevation model, we derive layers for slope and topographic wetness index. The point-location soil samples require more extensive processing, described below. Once this workflow is complete for each input, the covariate raster data are combined into a comma-separated text file that can be loaded in R. The input products are shown with spatial and temporal information in Table 1. The downloaded and processed input data are available in the downloadable package for this project.

| Name | Source | Spatial Resolution | Temporal Scale |
|---|---|---|---|
| gSSURGO | USDA | 10 m | variable |
| NCSS | USDA | point | 37 years (1960—1997) |
| NCDL | USDA | 30 m | 1 year (2015) |
| Aeroradiometrics | USGA | 2 km | 8 years (1973—1981) |
| GRIDMET | REACCH | 4 km | 30 years (1980—2010) |
| NED | USGS | 30 m | 83 years (1923—2016) |

Table 3.1 Datasets used as model inputs, including source, spatial resolution, and temporal scale

### *Study Area*

The REACCH study area is a polygonal area of about 93,800 km2 located within Northern Idaho, Eastern Washington, and Northeastern Oregon (Fig 3.1, left). This area encompasses the major cereal crop growing region of the inland northwestern United States. To minimize the occurrence of edge effects during statistical modeling, a bounding rectangle (122° W, 49° N, 115.5° W, 44° N) of approximately 277,000 km2 was constructed containing parts of Idaho, Montana, Oregon, and Washington (Fig. 3.1, right). The REACCH bounding rectangle is the spatial extent to which other data products are clipped during pre-processing.

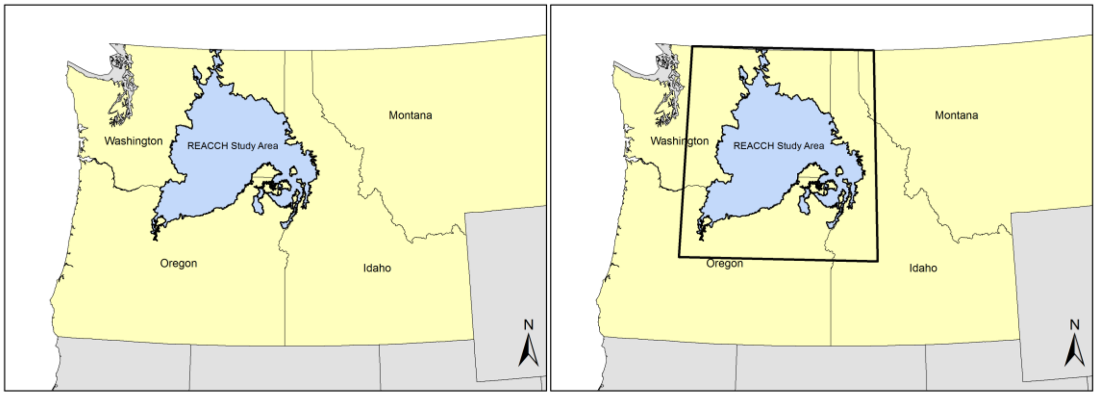

Figure 3.1 The Regional Approaches to Climate Change (REACCH) study area (left) and with bounding rectangle (right)

### *Carbon: USDA NCSS*

The USDA National Cooperative Soil Survey produces a Soil Characterization database that includes SOC (percent weight), bulk density (g/cm3), and percent rock content sampling at point locations throughout the US. Samples are taken per soil horizon, nominally for the full depth of the soil profile, data are not available for all horizons at every location. The NCSS SOC samples are used as ground-truthing data to train the random forest model.

Because the NCSS data are collected by disparate agencies across the study region, their form is more heterogeneous than other data used in the model and require more intensive processing to prepare for use. The data are downloaded as a set of comma-separated value (CSV) files for each county that intersects the REACCH bounding box. The CSV files for each county are joined using identifiers for each observation, and then values of interest (SOC percent weight, soil bulk density, and rock content) are extracted for each. Samples lacking any of the values of interest are discarded. Soil organic C content (g/m2) is computed following the method described by Bliss et al. (1995) for each soil layer and summed to find a total SOC value for each sample location. Sample locations outside the REACCH bounding box are removed and all remaining samples are added to a point-based shapefile. The process is repeated for each county, building upon the shapefile until all counties are processed.

### *Soil: USDA NRCS gSSURGO*

The USDA Natural Resources Conservation Service (NRCS) produces the Gridded Soil Survey Geographic Database, a 10 m spatial resolution gridded dataset with associated tabular data describing soil series and associated characteristics across the US. The gSSURGO dataset is based upon the SSURGO polygon vector dataset, a product of rasterizing the polygons. From the NRCS fact sheet on gSSURGO, "The raster map data have a 10-meter cell size that approximates the [SSURGO] vector polygons in an Albers Equal Area projection," which is to say, the vector polygons of SSURGO have been divided into 10 m grid cells and all the values of each polygon transferred to each grid cell (USDA NRCS, 2016). Because gSSURGO is based upon county soil survey data, the temporal scale varies depending on the update frequency of the counties. The associated tabular data include a measure of SOC (g/m2) and are used primarily as a basis for comparison of the model output of the scorpan process.

The data are downloaded as a spatial grid for each state (Idaho, Oregon, Washington) and are processed by mosaicking the state grids together, then extracting the grid cells from within the area of the REACCH bounding box, and finally the SOC value from the associated table is joined to the grid.

### *Climate: GRIDMET*

Given the importance of precipitation on SOC dynamics, a precipitation and temperature dataset has been included from Abatzoglou's Gridded Surface Meteorological Data (GRIDMET) (Abatzoglou, 2013). Mean annual temperature and mean annual precipitation have been shown to be significant predictors of SOC variability (Morrow, 2014). These data are 4 km spatial resolution raster data describing average precipitation and minimum and maximum temperature from 1979-2010. Using these data with a 4 km spatial resolution involves commission of the Ecological Fallacy, however, the importance of these climate variables on C dynamics in combination with the difficulty of obtaining higher resolution data make these data the best currently available for the purpose.

The data are downloaded from the Northwest Knowledge Network's (https://www.northwestknowledge.net) Thematic Real-time Environmental Distributed Data Services (THREDDS) server (http://www.unidata.ucar.edu/software/thredds/current/tds/), which allows the user to specify spatial and temporal bounds as well as aggregation criteria. Therefore, the downloaded data already represent the correct variables and spatial extent and require no pre-processing other than re-projection.

### *Organisms: USDA NASS NCDL*

The USDA National Agricultural Statistics Service (NASS) produces the National Cropland Data Layer (NCDL) yearly as a 30 m spatial resolution grid aligned to the 30 m National Elevation Dataset (NED) grid. The NCDL algorithmically classifies individual grid cells into agricultural cover types using satellite imagery, supervised classification techniques, and ground truthing. Since the area of interest of this study area is primarily agricultural land, the NCDL crop categories provide detailed information about land cover within the region.

The data are downloaded as a spatial grid for each state (Idaho, Oregon, Washington) and are processed by mosaicking the state grids together, then extracting the grid cells from within the area of the REACCH bounding box. Due to limitations of the random forest implementation in R, categorical inputs to random forest models are limited to a maximum of 53 classes (https://cran.r-project.org/web/packages/randomForest/NEWS). The clipped NCDL layer has 84 cover classes, which means that some classes must be collapsed (combined to form less specific classes). Care has been taken to avoid collapsing the classes of primary interest to cereal production and to prefer collapse of classes that are sparsely represented (or completely absent) within the REACCH study area. This aggregation of classes is done dynamically within the R code for the modeling; the cover classes stored on disk are left in their original form. Table 3.2 shows a mapping of classes that have been collapsed into more general groupings.

| New Class | Old Class | Pixel Count | Description |
|---|---|---|---|
| 44 | 6 | 86 | Sunflower |
| | 14 | 0 | Mint |
| | 44 | 6785 | Other Crops |
| | 58 | 2297 | Clover/Wildflowers |
| | 224 | 71 | Vetch |
| 47 | 41 | 773482 | Sugarbeets |
| | 47 | 1667 | Misc Vegs & Fruits |
| | 48 | 0 | Watermelons |
| | 50 | 0 | Cucumbers |
| | 54 | 0 | Tomatoes |
| | 55 | 0 | Caneberries |
| | 206 | 7040 | Carrots |
| | 207 | 0 | Asparagus |
| | 208 | 0 | Garlic |
| | 209 | 0 | Cantaloupes |
| | 216 | 541 | Peppers |
| | 219 | 477 | Greens |
| | 221 | 0 | Strawberries |
| | 227 | 2054 | Lettuce |
| | 242 | 0 | Blueberries |
| | 243 | 0 | Cabbage |
| | 246 | 2210 | Radishes |
| | 247 | 3937 | Turnips |
| 71 | 67 | 5265 | Peaches |
| | 70 | 710 | Christmas Trees |
| | 71 | 0 | Other Tree Crops |
| | 76 | 0 | Walnuts |
| | 77 | 311 | Pears |
| | 218 | 767 | Nectarines |
| | 220 | 681 | Plums |
| | 223 | 0 | Apricots |
| 121 | 121 | 2561272 | Developed/Open Space |
| | 122 | 1021456 | Developed/Low Intensity |
| | 123 | 394022 | Developed/Med Intensity |
| | 124 | 44875 | Developed/High Intensity |
| 222 | 222 | 0 | Squash |
| | 229 | 160 | Pumpkins |
| | 249 | 0 | Gourds |

Table 3.2 Land cover classification mapping

### *Relief: USGS NED*

The USGS National Elevation Dataset (NED) is a 30 m spatial resolution gridded digital elevation model (DEM) for the US. The varying topography of the study area influences erosional and depositional patterns of SOC across the landscape. Several products are derived from the NED: slope, a depression-filled DEM (O'Callaghan and Mark, 1984), flow accumulation and flow direction (Jenson and Domingue, 1988), and topographic wetness index (TWI) (Quinn et al., 1991; Moore et al., 1993; Gessler et al., 1995). Flow accumulation and direction are intermediate layers in this model; only elevation, slope, and TWI are used as model inputs.

The data are downloaded as a spatial grid for each state (Idaho, Oregon, Washington) and are processed by mosaicking the state grids together, then extracting the grid cells from within the area of the REACCH bounding box. Sinks are filled (Reuter et al., 2009) and a slope grid is generated, followed by flow direction, flow accumulation, and TWI.

### *Parent material and Age: USGS Aeroradiometric Grids*

The USGS aeroradiometric grids are 2 km spatial resolution maps of potassium, thorium, and uranium concentration in the top 30 cm of the Earth’s surface (Duval et al., 2005). The aeroradiometric data are thought to be a convenient proxy for parent materials (Bierwirth et al., 1996; Gessler et al., 1995). The data were collected between 1973 and 1981 by various contractors at various spatial resolutions, and were combined into a single dataset for the conterminous United States by the USGS with some data loss (Hill et al., 2009). Because of the heterogeneous composition of these data, their variable spatial and temporal resolutions, and their age relative to the rest of the input products, they may serve as a weak proxy for present-day parent material, but are nonetheless included as the best nationwide aeroradiometric product that currently exists. As with the climate data, the mismatch in spatial resolution between these data and other gridded inputs represents commission of the Ecological Fallacy.

The data are downloaded as a spatial grid for each element (potassium, thorium, uranium) for the continental US and are processed by extracting the grid cells from within the area of the REACCH bounding box.

### *Reproducibility*

To enable reproducibility, input data, metadata, computer code, and output products will be packaged together and made freely available for download. The input data for this model are publicly available datasets, however as new versions of datasets are released, the specific data used in this project may be eliminated, altered, or replaced. Therefore, input data are included in the packaging in their raw form, in addition to processed products of those input data.

Computer code that was used to download and/or perform ETL and pre-processing steps to prepare the input data for the modeling process is also included. The ETL programs developed are able to collect data from the necessary repositories, process the data, and generate metadata that describes the data processing steps used to prepare data for model ingestion. For example, the NED data are used to derive other products such as slope and TWI, which then go on to be inputs to the modeling process. These intermediate, processed data are also included in the project package, in order to demonstrate that the ETL code verifiably produces the intermediate data, and that the intermediate data are then used as inputs to the modeling process.

The R code that implements the modeling process itself is also included. The random forest process depends upon a pseudo-random number generator (PRNG) to execute. Owing to different implementations of PRNG, and of floating point operations and representation across different computer architectures, model outputs may not be exactly identical to the output produced by this project. However, the architecture and software used to produce these outputs are documented in the project metadata files, and a random number seed is fixed within the model code in order to maximize the replicability of the exact model outputs.

The publishable results of this project will be the product layers, the ETL and model programs used to create the product layers, appropriate metadata documentation, and this paper describing the development and implementation of the software and product layers. All products are freely available in an on-line repository and licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0) (https://creativecommons.org/licenses/by-sa/4.0/legalcode). In short, this license expresses that others who wish to use any part of this project are free to do so in any context they wish, so long as they agree to provide attribution to the authors in any publication they make, and agree to make their derived products likewise accessible.

## Analysis

McSweeney et al. (1994) provide a foundation for modeling soil characteristics using GIS and soil horizon characteristics. We combine this approach with the scorpan model, formalized by McBratney et al. (2003), which describes a set of covariates used as inputs to digital soil models. These inputs are soil, climate, organisms, topography, parent material, age, and spatial position. Odgers et al. (2013) applied the scorpan model in combination with an algorithm called DSMART—Disaggregation and Harmonisation of Soil Map Units Through Resampled Classification Trees. Chaney et al. (2016) extended the DSMART algorithm to work in supercomputing environments (DSMART-HPC) and applied the scorpan method to develop a soil series map of the contiguous United States. As an example analysis of the environmental data collected, and following the general methods of these papers, we develop a model using scorpan inputs to develop a map of SOC. Because soil C is a continuous variable, a classification algorithm is not an appropriate predictive tool, and a random forest regression algorithm is substituted. The scorpan approach explicitly supports the use of modeling continuous attributes of soil (McBratney et al., 2003). With the exception of the point-based measurements of SOC used to train the model, all of the explanatory variables take the form of geospatial gridded datasets.

### *Model Selection, Performance, and Diagnostics*

A model selection process was followed to choose up to seven input variables (this limit was chosen as the maximum based upon memory constraints). Each of mean annual temperature, mean annual precipitation, NCDL classification, elevation, slope, TWI, potassium, thorium, and uranium were introduced into the model, replacing the least important terms when the limit was reached. By this method, a random forest model was specified to predict the natural log of SOC for the full available soil depth using elevation, thorium (Th), uranium (U), NCDL classification, slope, mean annual temperature, and mean annual precipitation:

$$\ln(\text{SOC}) \sim \text{elevation}+\text{Th}+\text{U}+\text{NCDL}+\text{slope}+\text{temperature}+\text{precipitation}$$

Table 3 shows these variables in order of importance measured as the increase in mean squared error (MSE) when their values are randomly permuted during model training. The size of the random forest was fixed at 512 trees, based upon Oshiro et al. (2012) showing that increasing the number of trees far beyond 128 is likely to show little benefit for the increase in processing time due to the low number of covariates in the model leading to asymptotically small performance gains of additional trees; as processor speed has improved since 2012, this threshold was multiplied. The random forest approach produces a pseudo-R2 value that is equal to 1-(MSE/Variance), and it indicates the model's benefit over a null model of using the grand mean of the independent variable as the prediction. This model produces a pseudo-R2 of 20.49%, which is acceptable for the proof-of-concept purpose of this paper. The model mean squared residuals value of 0.25 is low, though the value is reported in log-transformed units and may represent overfitting of the model. Figure 3.2 is a map of predicted SOC for the REACCH study area. Figure 3.3 is a map of the same area using SSURGO SOC values, for comparison. In both maps, a general West to East increase in SOC within the REACCH study area is shown. An East-West precipitation gradient over the area is a primary SOC driver (Morrow, 2014), reflecting the importance of precipitation in the model. The maps are particularly different in the southeastern portion of the study area within Idaho, where very few training sample locations were found. Figure 3.4 is a map of the variance of the random forest estimator for each point on the grid. The relatively high variance in that same area of Idaho indicates less stability in the random forest estimator in that area.

| **Explanatory Variable** | **% Increase in MSE** |
|---|---:|
| Mean Annual Precipitation | 22.16 |
| Mean Annual Temperature | 15.61 |
| Elevation | 12.94 |
| Slope | 12.29 |
| NCDL Class | 11.21 |
| Thorium | 9.33 |
| Uranium | 7.79 |

Table 3.3 Results of variable selection

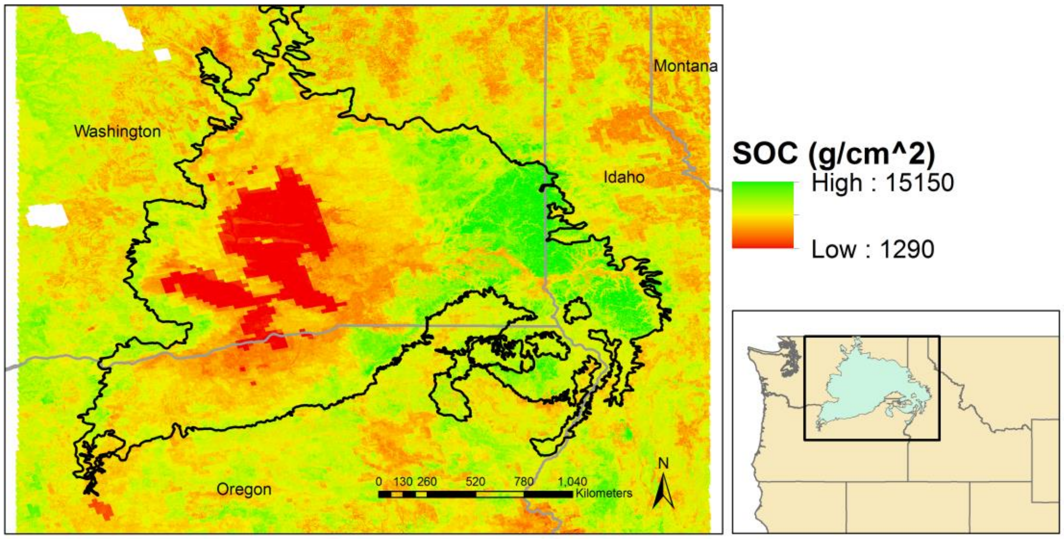

Figure 3.3 The random forest soil organic carbon map

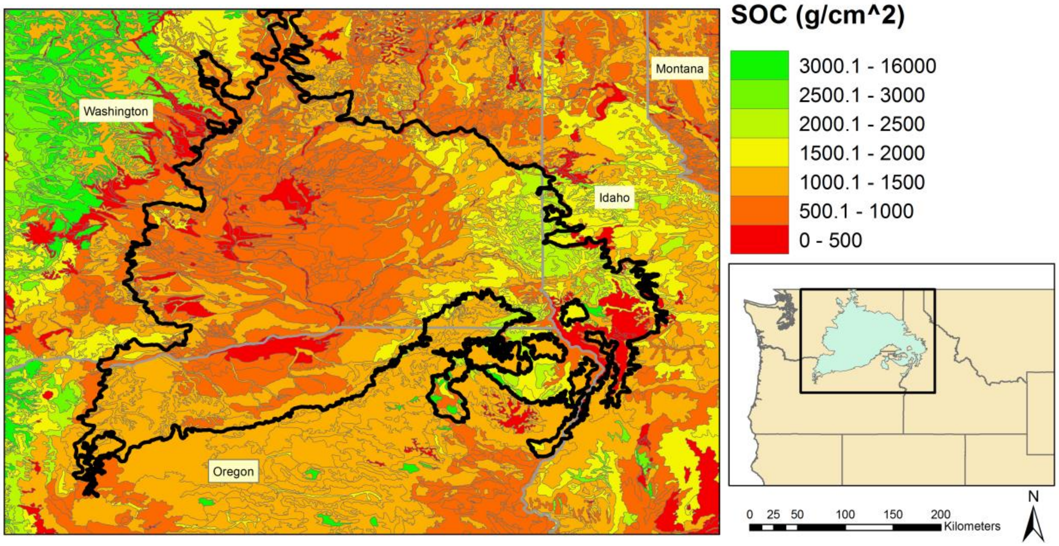

Figure 3.2 The Soil Survey geographic (GSSURGO) soil organic carbon map for comparison

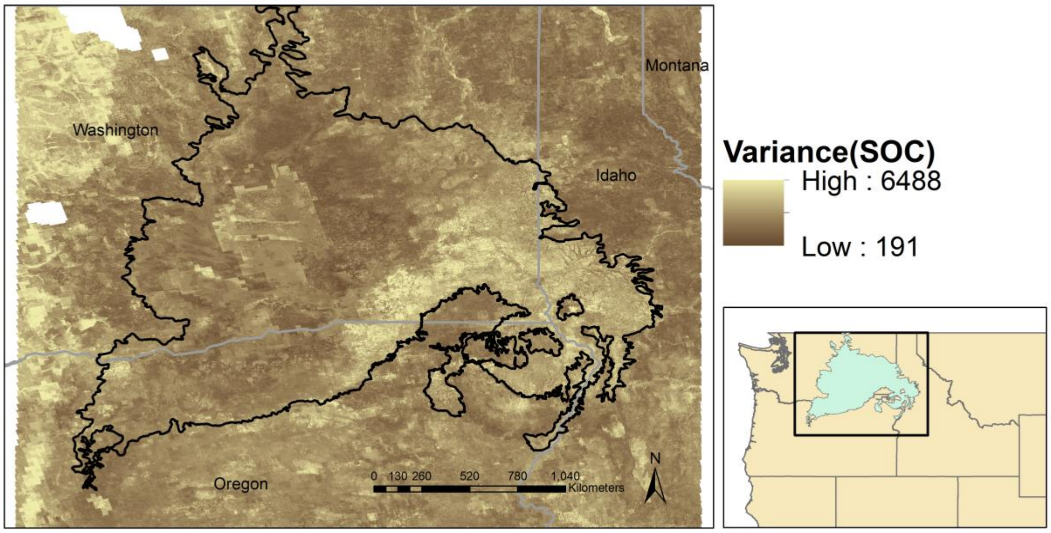

Figure 3.4 Variance map of the soil organic carbon estimator

## Discussion

### *Covariates*

The pool of covariates was chosen because of their historical inclusion in similar modeling exercises (Jenny, 1941; McBratney et al., 2003; Morrow, 2014), and an exhaustive effort to identify additional or different covariates was not undertaken. The primary goal of this paper is to develop and describe a repeatable framework as a first step towards a more accurate SOC model. It would be informative to repeat the modeling process with a different land cover classification product such as the National Land Cover Database. Additional covariates could also be added, including Multiresolution Valley Bottom Flatness Index (MRVBF) (Gallant and Dowling, 2003); geological data that could help characterize the parent material at a spatial resolution that is better or meets the 30 m resolution of the other covariate layers; and other covariates as needed. The publication of the input data and processing code required to execute the model make it feasible for the community to further evolve the model with new and different covariates or different modeling approaches. The intent of the framework approach is, at least initially, to produce a foundation for development and comparison of approaches to SOC modeling (and potentially other soil- or agriculture-centric climate analyses).

### *The Ecological Fallacy*

The 4 km grid cell footprint of the climatic variables and the 2 km grid of the USGS aeroradiometric data appears prominently in areas throughout the modeled output product. This is a consequence of the assumption that the values of these large grid cells are representative of the values of the 30 m cells that are overlaid from the other gridded products. This assumption is an example of the Ecological Fallacy, in which a characteristic of aggregate data is simply applied to groups within the aggregate (Selvin, 1958; Piantadosi, 1988).

Two approaches to eliminating this issue from the result are to aggregate the other input layers from 30 m up to 4 km, thus creating a complete set of inputs that are comparable in spatial scale; or to use an alternate data layers for climate and aeroradiometrics (possibly even derived from the current layers using a statistically valid downscaling technique) to provide influence from climate and parent material in the model. The latter approach is beyond the scope of this paper, and the former approach was not chosen in order to highlight the existence of this common issue in spatial modeling and to illustrate the effects of a third choice: using the data as available and explicitly noting the incompatibility and the existence of statistical flaws in the approach. Since much of the spatial data that we use are gathered from external providers like USGS, there are limitations to the compatibility of our various layers. In some cases, we may find that our models work best with data that are incompatible in some ways, and it is important to advertise these incompatibilities to readers and potential users of our products.

### *Spatial Extent and Scale*

The spatial extent of this modeling effort was driven by the boundaries of the area of interest to the REACCH-PNA project. Given that the boundaries were established based upon cereal production capacities and practices in the region, there is reduced variability within some or all of the variables of the model that suggest the suitability of this specific model and its inputs may be significantly different when applied to different geographic areas. Nonetheless, the general applicability of the scorpan model and the relatively nonspecific implementation framework presented here could be readily adapted to other spatial extents.

The 30 m spatial resolution of the input and output products was chosen as a result of the availability of input products at that resolution, driven primarily by the USGS elevation grid, which has been adopted by the creators of other gridded datasets to allow for convenient overlay. Depending on the spatial scale of physical processes involved in the model and the intended application of any output products of the modeling process, the 30 m grid may or may not be an appropriate resolution. The USGS also makes available a 10 m gridded elevation layer, which may be more suited to certain types of analysis; however, it may be difficult or impossible to assemble necessary covariate layers at that resolution. Again, the code framework described here is readily adapted to process data of various spatial resolutions.

### *Temporal Considerations*

One challenge of building these types of models is the availability of data products that are collected at temporal scales similar to their natural variability. Some of the covariate layers, such as elevation, are not likely to change dramatically over short periods of time. Other layers, such as land cover classification, may change significantly from year to year, particularly in agricultural areas where cropping systems drive the rotation of different crops into fields over time, and where crop selection may be market-driven. The NCSS soil samples have been taken over a range of years that is not necessarily expressed in the data tables, and while some areas may experience relatively stable soil conditions, this is not necessarily the case in agricultural areas under various management regimes. The mean annual temperature and precipitation layers are the two most important layers to the model, and are also of great interest considering changes expected under various likely regimes of climate change. It is unclear how much influence the variability over time of some input products have on the model output, but assessing the stability of the modeling process over time could be informative.

## Summary

The reproducibility crisis is spreading in the sciences, and in light of its inherent complications, particularly with respect to climate change research, it is important that researchers embrace open science principles. Science is a process of building upon existing work, and by publishing all components of research including input, processing code, and output, we make explicit the foundation upon which new science can be built.

To that end, this paper describes a framework of input data, processing code, and outputs designed around the concept of modeling SOC for the cereal grains producing region of the Northwestern United States. The framework can be improved iteratively with updated versions of its existing covariates and with new covariates as they become available, with alternate data processing tools, and with improved statistical models. Run periodically for covariates representing different time periods, the framework could be used to model C dynamics over time. The framework can furthermore be altered to focus on different geographic areas or scales and to model other environmental variables, related to soil or otherwise. The flexibility and re-usability of the framework makes it a potential foundation for more extensive modeling efforts, but it also makes explicit the processes that have gone in to producing its results.

As we address reproducibility and the rapid pace of modern science, we can help ourselves by embracing open science:

publishing input data, ensuring that older versions of the data remain available and uniquely identifiable to preserve replicability

publishing computer code, ideally in languages, APIs, and tools that are freely available, and ideally complete enough to replicate published results with relative ease

publishing processed data that allows others to cross-check both processing code, processed outputs, and result data

publishing a paper that traditionally describes research motives, methods and results

applying an explicit license to all published products to ensure that downstream users are aware of their rights and responsibilities when using data or code

creating persistent identifiers that make it easier for downstream users to verify that they are using the correct data and code

using embargo periods that are long enough to avoid getting scooped by other researchers, but not so long that the data and code are obsolete by the time the embargo expires

The soil C modeling project described in this paper has followed these steps, and the input data, computer code, and output data are all accessible using this DOI: 10.7923/G4XP72ZB. The computer code is also available on GitHub at https://github.com/flathers/soilCarbonFramework.

## References

Abatzoglou, J.T.  2013.  Development of gridded surface meteorological data for ecological applications and modelling.  International Journal of Climatology 33.1:121–131. doi:10.1002/joc.3413

Asendorpf , J.B., M. Conner, F. De Fruyt, J. De Houwer, J.J.A. Denissen, K. Fiedler, et al.  2013. Recommendations for increasing replicability in psychology.  European Journal of Personality 27.2:108–119.  doi:10.1002/per.1919

Baker, M.  2016.  Is there a reproducibility crisis?  Nature 533.7604: 452–454.  doi:10.1038/533452a

Bierwirth, P., Gessler, P., and McKane, D. 1996. Empirical investigation of airborne gamma-ray images as an indicator of soil properties Wagga Wagga, NSW. AGSO record (1996).

Bliss, N. B., S.W. Waltman, and G.W. Petersen. 1995. Preparing a soil carbon inventory for the United States using geographic information systems. Soils and global change (1995): 275-295.

Chaney, N.W., E.F. Wood, A.B. McBratney, J.W. Hempel, T.W. Nauman, C.W. Brungard, and N.P. Odgers.  2016.  POLARIS: a 30-meter probabilistic soil series map of the contiguous United States.  Geoderma 274:54–67.  doi:10.1016/j.geoderma.2016.03.025

Cragin, M.H., C.L. Palmer, J.R. Carlson, and M. Witt.  2010.  Data sharing, small science and institutional repositories.  Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 368.1926:4023-4038. doi:10.1098/rsta.2010.0165

Doetterl, S., A.A. Berhe, E. Nadeu, Z. Wang, M. Sommer, and P. Fiener.  2016.  Erosion, deposition and soil carbon: a review of process-level controls, experimental tools and models to address C cycling in dynamic landscapes.  Earth-Science Reviews 154: 102-122. doi:10.1016/j.earscirev.2015.12.005

Duval, J.S., J.M. Carson, P.B. Holman, and A.G. Darnley.  2005.  Terrestrial radioactivity and gamma-ray exposure in the United States and Canada: U.S. Geological Survey Open-File Report 2005-1413.  http://pubs.usgs.gov/of/2005/1413/ (Accessed 8 Nov. 2016).

Eklund, A., T.E. Nichold, and H. Knutsson.  2016.  Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates.  Proc. Natl. Acad. Sci. U. S. A. 201602413. doi:10.1073/pnas.1602413113

FAO. 2017. Global Soil Organic Carbon Map. http://www.fao.org/global-soil-partnership/pillars-action/4-information-and-data/global-soil-organic-carbon-gsoc-map/en/ (accessed 12 Nov. 2017).

Florinsky, I. V. 2012. The Dokuchaev hypothesis as a basis for predictive digital soil mapping (on the 125th anniversary of its publication). Eurasian soil science 45.4: 445. doi: 10.1134/S1064229312040047

Gallant, J.C., and T.I. Dowling. 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resources Research 39, no. 12. doi:10.1029/2002WR001426

Gessler, P.E., I.D. Moore, N.J. McKenzie, and P.J. Ryan. 1995. Soil-landscape modeling and spatial prediction of soil attributes. Special issue: Integrating GIS and Environmental Modeling. International Journal of Geographical Information Systems, Volume 9(4):421-432.

Gessler, P. E., O.A. Chadwick, F. Chamran, L. Althouse, and K. Holmes. 2000. Modeling soil–landscape and ecosystem properties using terrain attributes. Soil Sci. Soc. Am. J. 64.6:2046-2056. doi:10.2136/sssaj2000.6462046x

Gezelter, J.D. 2015. Open source and open data should be standard practices. The journal of physical chemistry letters 6.7:1168–1169. doi:10.1021/acs.jpclett.5b00285

Hill, P.L., R.P. Kucks, and D. Ravat. 2009. Aeromagnetic and aeroradiometric data for the conterminous United States and Alaska from the National Uranium Resources Evaluation (NURE) Program of the U.S. Department of Energy: U.S. Geological Survey Open-File Report 2009–1129. http://pubs.usgs.gov/of/2009/1129/index.html (Accessed 8 Nov. 2016).

Jenny, H. 1941. Factors of Soil Formation, A System of Quantitative Pedology. McGraw-Hill, New York.

Jenson, S.K., and J.O. Domingue. 1988. Extracting topographic structure from digital elevation data for geographic information system analysis. Photogrammetric engineering and remote sensing 54, no. 11:1593-1600. doi:10.1.1.138.6487

Kaisler, S., F. Armour, J. A. Espinosa, and W. Money. 2013. Big data: Issues and challenges moving forward. Proceedings of the Hawaii International Conference on System Sciences. Hawaii. doi:10.1109/HICSS.2013.645

Lin, J. W. B. 2012. Why Python is the next wave in earth sciences computing. Bulletin of the American Meteorological Society, 93(12), 1823-1824. doi:10.1175/BAMS-D-12-00148.1

Marshall, E.  2002.  DNA sequencer protests being scooped with his own data.  Science 295.5558:1206-1207.  doi:10.1126/science.295.5558.1206

McBratney, A.B., M.L. Mendonca Santos, and B. Minasnya.  2003.  On digital soil mapping.  Geoderma 117.1:3–52.  doi:10.1016/S0016-7061(03)00223-4

McNutt, M.  2012.  Reproducibility.  Science 343 (6168), 229. doi:10.1126/science.1250475

McSweeney, K., B.K. Slater, R.D. Hammer, J.C. Bell, P.E. Gessler, and G.W. Petersen. 1994. Towards a New Framework for Modeling the Soil-Landscape Continuum. In: R.R. Amundson, J. Harden, M. Singer, editors, Factors of Soil Formation: A Fiftieth Anniversary Retrospective, SSSA Spec. Publ. 33. SSSA, Madison, WI. p. 127-145. doi:10.2136/sssaspecpub33.c8

Moore, I.D., P.E. Gessler, G.A. Neilsen, and G.A. Petersen. 1993. Soil attribute prediction using terrain analysis. Soil Science Society of America Journal. 57:443-452.

Morrow, J.G. 2014. The influence of climate and management on surface soil health within the inland Pacific Northwest (Master's thesis). Retrieved from http://www.dissertations.wsu.edu/Thesis/Summer2014/j_morrow_071414.pdf

O'Callaghan, J.F., and D.M. Mark.  1984.  The extraction of drainage networks from digital elevation data.  Computer vision, graphics, and image processing 28, no. 3:323-344. doi:10.1016/S0734-189X(84)80011-0

Odgers, N. P., W. Sun, A. McBratney, B. Minasny, and D. Clifford.  2014.  DSMART: An algorithm to spatially disaggregate soil map units. Geoderma 214: 91–100. doi:10.1016/j.geoderma.2013.09.024

Open Science Collaboration.  2015.  Estimating the reproducibility of psychological science. Science 349, aac4716.  doi:10.1126/science.aac4716

Piantadosi, S., D.P. Byar, and S.B. Green.  1988.  The ecological fallacy.  American Journal of Epidemiology 127(5), 893-904.

Oshiro, T.M., P.S. Perez, and J.A. Baranauskas.  2012.  How many trees in a random forest?  In: Perner, P., editor, Machine Learning and Data Mining in Pattern Recognition.  8th International Conference, MLDM 2012, Berlin, Germany.  13–20 Jul. 2012.  Springer Berlin Heidelberg, Berlin, Heidelberg.  doi:10.1007/978-3-642-31537-4_13

Quinn, P., K. Beven, P. Chevallier, and O. Planchon. 1991. The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. Hydrol. processes 5.1:59–79. doi:10.1002/hyp.3360050106

Raich, J. W., and W. H. Schlesinger. 1992. The global carbon dioxide flux in soil respiration and its relationship to vegetation and climate. Tellus B 44.2:81–99. doi:10.1034/j.1600-0889.1992.t01-1-00001.x

Reuter, H.I., Hengl, T., Gessler, P., and Soille, P. Preparation of DEMs for geomorphometric analysis. 2009. Developments in Soil Science 33: 87-120.

Selvin, H.C. 1958. Durkheim's suicide and problems of empirical research. American journal of sociology 63.6:607–619. doi:10.1086/222356

Stockle, C.O., M. Donatelli, and R. Nelson. 2003. CropSyst, a cropping systems simulation model. Eur. J. Agron. 18.3:289–307. doi:10.1016/S1161-0301(02)00109-0

USDA NRCS. 2016. Gridded soil survey geographic (gSSURGO) database fact sheet. USDA NRCS. https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs142p2_052164.pdf (Accessed 8 Nov. 2016).

Vassiliadis, Panos. 2009. A survey of extract–transform–load technology. International Journal of Data Warehousing and Mining 5.3:1–27. doi:10.4018/jdwm.2009070101

# Chapter 4: Methods for Expressing Machine-Readable License Information in Geospatial Metadata

## Introduction

As the paradigm of Open Science has developed, researchers have been working to define exactly what the term implies. The Panton Principles for Open Data in Science (https://pantonprinciples.org/) provide a starting point for developing data sharing practices (Murray-Rust et al., 2010). The FAIR Principles for data management and stewardship further develop the specific requirements of “open” science and data to include the concepts of findability, accessibility, interoperability, and reusability (Wilkinson et al., 2016). At its foundation, Open Science is about sharing, and successful sharing implies common understanding between involved parties of behavioral rules and boundaries. In the information era, data can easily be made available for consumption via Internet-based services including application program interfaces (APIs), web-based applications, and simple exposure of files on web servers, among other methods. Data can also be shared via less formally defined methods—as attachments to email or through file sharing services such as DropBox, for example.

The results of a survey of scientists across a variety of disciplines published in 2015 show that data sharing is becoming more common—researchers are increasingly making their data available to others—and by methods that are easier for data consumers to access (Tenopir et al., 2015). The survey data also show that public research funding agencies are increasingly requiring researchers to share data resulting from funded projects (Tenopir et al., 2015). As technologies for data sharing grow easier to use and attitudes shift in favor of sharing research data, it becomes more likely that consumers of research data may have no relationship with data creators. Where in the past, data may have been passed from lab to lab by colleagues, today it is more often accessed through a web page.

When data sharing was more commonly a person-to-person activity, it was easy to have conversations about appropriate use of the data. Creators could pass on warnings along with data, for example if data were restricted from publication due to a legal agreement with the contractor that provided the data. Researchers who wished to re-use a colleague’s data to support a publication could discuss appropriate ways to assign credit to data creators. Today, in a more impersonal data sharing environment, these conversations between creators and consumers can be more difficult for various reasons, including the geographic distance between them, language barriers, and others.

As data sharing transitions away from informal interpersonal arrangements toward impersonal agreements, more formal language becomes appropriate to define those agreements. In order for data creators to publish data with confidence that they are not opening themselves to legal liability—for example, from data consumers who misuse data for purposes to which they are not suited—creators should include liability limitation statements with their data.

Likewise, data consumers need to be aware of their rights—and any limitations on their rights—when using data they have acquired from creators. Consumers can only be assured of their rights when data are accompanied with formal language describing the rights and restrictions granted to consumers of the data.

The formal language required to express concepts like limitation of liability, rights, and restrictions, is legal language. These collections of legal concepts are referred to as "licenses." In terms of the FAIR Principles, data reusability demands that "data are released with a clear and accessible data usage license" (Wilkinson et al., 2016). Since most researchers are not lawyers, it is generally unwise for them to attempt to develop the text of licenses themselves. Fortunately, there exist a number of licenses that have been developed in the Internet era that are freely available for content creators to apply to their works upon distribution. Using pre-existing licenses in the distribution of work helps to reduce the proliferation of novel licenses, reducing the amount of overhead required for both creators and consumers to understand the specific features of licenses with which they engage (Katz, 2006). The Free and Open Source Software community has a long history of defining and applying rules for sharing program code through licenses, among them being the Creative Commons (CC), BSD 3-Clause License (BSD-3-Clause), and the GNU General Public License version 3.0 (GPL-3.0). These licenses have been developed primarily to apply to software, but may be applicable to research data, in some cases (Stodden, 2008).

One limitation of the licenses available is that they are based upon copyright law, and in some jurisdictions (such as the United States), data are not eligible for copyright protection because they are not considered to be creative works. In those cases, there is still existing language that creators may use when publishing data, though these approaches lack some of the features allowed in copyright-based licenses. In other (particularly European) countries, data collections may be eligible for copyright protection based upon sui generis provisions designed to recognize the investment of the creator or collator in a data collection (Guibault, 2013; Khayyat & Bannister, 2015). Some existing licenses explicitly recognize the sui generis provisions, where they exist.

Legal considerations of data licensing vary widely by jurisdiction and over time as legal requirements change (Bedini et al., 2014; Guibault, 2013; Khayyat & Bannister, 2015; Korn & Oppenheim, 2011; Lee, Allard, McGovern, & Bishop, 2016) and are largely beyond the scope of this paper. Though it is not the purpose of this paper to offer legal advice, and the authors are not lawyers, an overview of several licenses and other approaches to formalizing sharing agreements will be undertaken.

In the era of data-intensive science moving towards an ideal, "in which all of the science literature is online, all of the science data is online, and they interoperate with each other," simply associating sharing agreements with data is not sufficient (Hey 2009). To enable large-scale, data-intensive research projects, sharing agreements must be attached to data in standard ways and expressed in standard language that allows computer systems to locate and ingest the sharing agreements attached to a given dataset. In the interoperability section of the FAIR Principles, this need is expressed as a requirement that "data use a formal, accessible, shared, and broadly applicable language for knowledge representation" (Wilkinson et al., 2016). In research projects that make use of potentially vast numbers of datasets collected from a variety of creators, the only way to keep adherence to sharing agreements manageable is if they are standardized to make them machine-readable.

In addition, machine-readable sharing agreements enable researchers to streamline other aspects of data-intensive projects. As data become available through web portals and other aggregation services, users can search for data based upon their licensing requirements—allowing them to filter out data that are incompatible with the intended application. Automated systems such as extract, transform, load (ETL) processes could be programmed to make decisions about the incorporation of data based upon the machine-readable sharing agreements. Prior to the release or publication of products based upon large-scale integration of external data, automated audits could be executed to verify that all included data are allowed to be included in the final product. Further, as Contreras and Reichman point out, "legal interoperability can enable researchers to access and use data across multiple repositories without seeking authorization on a case-by-case basis, which increases the likelihood that more data will be put to productive use" (Contreras & Reichman, 2015). Repository and data managers can also use these machine-readable agreements to scale up repository management actions and identify problematic or challenging data collections for further review or treatment.

If data sharing agreements must be attached to data in a standard way, one direct approach is to encapsulate agreements within metadata records that accompany datasets: “Whenever possible, use metadata to indicate the licensing terms explicitly,” (Williams et al., 2014). Metadata come in a variety of standards, but according to the survey by Tenopir et al. (2015), the most common types of metadata included with research data follow the Dublin Core (DC), the Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSGDM), Ecological Metadata Language (EML), and the International Standards Organization (ISO) geographic information standard ISO 19115. Each of these standards allows (or can be adapted to allow) for the inclusion of information describing the sharing agreement under which the data are published. They also allow for an independent expression of the sharing agreement under which the metadata, themselves, are published, which is especially important when metadata are published independently of related data.

Standard expressions of sharing agreements can be more flexible, depending upon community behavior and needs. While all of the sharing agreements discussed in this paper can be expressed as full (or abbreviated) English-language texts, there also exist succinct and intentionally machine-readable expressions. Full-text versions of licenses can be considered to be machine-readable, as long as they use the same standard language every time they are expressed. Difficulties can arise, however, when these texts include white space and punctuation that may or may not be included in a particular instantiation of the agreement. Where possible, it can be more convenient to use very short texts that encode terms for sharing in unambiguous machine-readable language.

The Creative Commons provides two ways of expressing sharing agreements as machine-readable text: by using the Creative Commons Rights Expression Language (CC REL), or by crafting a Uniform Resource Identifier (URI) that contains the basic features of the agreement as well as providing a link to the full version of the agreement stored on the Creative Commons web site. The CC REL is most conveniently used to express CC licenses, and may be more challenging to apply to licenses outside the CC realm (fixme: citation).

The Software Package Data Exchange (SPDX) does not produce sharing agreements, but catalogs them: they provide a list of more than 200 of the most common sharing agreements and standardized expressions for each (Odence et al., 2015). The list includes a unique identifier and a selection of machine-readable expressions for each sharing agreement. The number of licenses represented by the SPDX list combined with their ease of identification and expression makes the list an ideal resource for embedding sharing agreements in metadata records; in fact, the EML Project has chosen to specifically recommend SPDX identifiers for sharing agreements in the latest version of their metadata standard (Jones et al., 2019)

A license can be referenced from the SPDX list using JavaScript Object Notation (JSON) encoding. The JSON text may include a number of fields, including the license text, name, unique identifier, and a link to the license source. A second way of referencing a license from the list is using a simple SPDX ID, which is a simple text line that contains the string “SPDX-License-Identifier:” followed by the unique identifier of a license from the list (Software License Data Exchange, 2022).

We perform a review of sharing agreements including licenses and waivers, several metadata standards’ support for rights expression, and suggest a common strategy for expressing data sharing agreements in machine- and human-readable formats encapsulated within standard metadata documents.

## Review of Sharing Agreements

Some popular licenses used in open source software are the Apache License 2.0 (Apache-2.0), Creative Commons (CC) family licenses, BSD 3-Clause "New" or "Revised" License (BSD-3-Clause), GNU General Public License version 3.0 (GPL-3.0), and the MIT license (MIT). Each of these licenses has a long form of the text of the license that could be embedded within a free text field in an XML metadata, but some run to thousands of words and contain formatting that may not be easily maintained within an XML structure. Each also has an SPDX JSON expression that is more conveniently applied in an XML context.

The common foundation of all of these licenses is copyright law. All of these licenses originated in the United States and therefore may be particularly associated with the specifics of US copyright law. Particularly with the CC licenses, there have been efforts to integrate the licenses with common legal requirements of other countries (González, 2015).

Because licenses are founded in copyright law, they may not be appropriate instruments for applying sharing agreements to data. The precise nature and extent of these circumstances are beyond the scope of this paper. Arguments about whether data are eligible for copyright and whether open source licenses are enforceable are ongoing in various jurisdictions and will continue to be decided by courts and governing bodies (Gomulkiewicz, 2011).

A more widely applicable approach to defining sharing agreements for open data is an exception or waiver of rights automatically granted under applicable intellectual property laws. This can be somewhat complicated by the variety of rights granted in different jurisdictions, but care has been taken in the development of formal waivers to recognize the most common situations. In some jurisdictions, such as France and Germany, there exist rights that cannot be entirely waived, such as moral rights (one example of a moral right is the right to attribution, that is, acknowledgement of authorship) (Sundara Rajan, 2011, p. 68). Nonetheless, the waivers attempt to clearly express the intent of the data developer in applying these types of sharing agreements.

In the US, some works are ineligible for copyright protection and are required to be released into the public domain, particularly those generated by the federal government and employees (Copyright Law of the United States, 2016). Again, there are complexities in whether data produced by or for various agencies within the government are necessarily released into the public domain, and this paper makes no assertion or recommendation regarding the legal status of data produced by or for the US federal government. However, given that a dataset is to be released to the public domain, a waiver may be an appropriate sharing agreement for clarifying the status of the data.

The Creative Commons offers two waivers that may be appropriate sharing agreements for information in the public domain. One is the Public Domain Mark; the other is Creative Commons Zero (CC0). Creative Commons recommends the Public Domain Mark for material that are unambiguously in the public domain in all jurisdictions, usually due to extreme age. They specifically recommend against using the Mark for material that may be encumbered by copyright restrictions in some jurisdictions and not in others.

The CC0 waiver is an attempt at a sharing agreement that enables a creator to formally waive all copyright rights to the extent possible (in some jurisdictions, there may remain rights that an individual is not legally able to waive). This agreement is appropriate for a creator who wishes to disclaim rights to a data product.

## Review of geospatial metadata standards

In addition to the four most common metadata standards in use (DC, FGDC, EML, and ISO 19115), we have included a more recent version of the ISO standard, ISO 19115-3 for comparison.

The primary (and, often, only) language of expression of these metadata standards is Extensible Markup Language (XML). Each of the metadata standards treated here has an expression in XML that can be defined by a schema document that exactly specifies the structure of a metadata document and, at least syntactically (though not semantically), the content that is allowed within elements of that structure. Schema documents are typically provided by the creator or maintainer of the metadata standard. XML is an ideal language for expressing machine-readable payloads in part because of the structure provided by the schema, which allows a computer to parse a metadata document and associate specific elements with specific meanings.

The XML schema for each metadata standard provides a free-text field for expressing use restrictions. There are several limitations of free-text within XML, particularly characters that are “reserved” for special meaning within the XML language. These characters are ampersand (&), left angle bracket (<), and right angle bracket (>). The reserved characters can still be represented in XML; for example, the ampersand can be encoded as “&amp;” (incidentally demonstrating why the ampersand itself is a reserved character in the language—it is used as a marker for the start of encoding of a special character). In general, other printable Unicode characters are allowed. If machine-readable license information can be expressed without the use of the prohibited characters, the elements will support the information readily. In cases where the reserved characters are needed in the license expression syntax, it would be necessary to use encoded versions in the XML.

In case forbidden characters are needed or desirable, there is a method for embedding text within an XML document that will not be interpreted by the parser. The Character Data (CDATA) markup block can be embedded within an XML element and is allowed to include the normally prohibited characters. Although CDATA blocks can be used to embed otherwise forbidden XML expressions of sharing agreements in existing standards-based metadata, in practice it can be simpler to embed sharing agreements using forms of expression that do not conflict with the XML environment. JavaScript Object Notation (JSON) is a format for encoding data that can be used to express a sharing agreement as a set of key/value pairs. Because JSON notation does not rely on the use of any of the reserved characters of XML, JSON data fragments can be embedded in an XML metadata in any free-text field without the need for CDATA blocks and without violation of the metadata schema.

## Placement of Sharing Agreement Information in Specific Metadata Standards

For each standard, there is an excerpt of XML code showing the specific XML structure needed to support the element (Figures 1-8). The validation of each example XML metadata as well-formed and compliant with the corresponding schema was performed using the Oxygen XML Editor (https://www.oxygenxml.com/).

### *Sharing Agreements Dublin Core Elements and Metadata Terms*

Dublin Core (DC) metadata can come in several forms, owing to the history of the development of the standards. The simplest form of DC emerged from a workshop held by the Online Computer Library Center (OCLC) and the National Center for Supercomputing Applications (NCSA) and is today known as Dublin Core Elements. It is a collection of 15 elements that have been recognized as ANSI and ISO standards. This form of DC metadata is intentionally very flexible in its implementation and content. The expressivity of DC has been both duplicated and extended with additional terms, published as an updated ISO standard in 2019. The DCMI today recommends using the newer DC Terms standard over the older Elements.

In the simple Elements form of DC, a free-text field called “rights” is intended to contain any information about intellectual property rights. With no constraints on the content of the field, there is no difficulty in expressing a sharing agreement as SPDX JSON within the rights field.

```
<?xml version="1.0" encoding="utf-8"?>
<metadata
xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance
xsi:noNamespaceSchemaLocation="http://purl.org/dc/elements/1.1/">
  [...]
  <rights>
      {
      "isDeprecatedLicenseId": false,
      "isFsfLibre": true,
      "licenseText": "Creative Commons Legal Code\n\nCC0 1.0 [...]",
      "standardLicenseTemplate": "[...]",
      "name": "Creative Commons Zero v1.0 Universal",
      "licenseId": "CC0-1.0",
      "seeAlso": [
        "https://creativecommons.org/publicdomain/zero/1.0/legalcode"
      ],
      "isOsiApproved": false
      }
  </rights>
</metadata>
```

Figure 4.1 Sharing agreement expression in Dublin Core Elements

The DC Terms standard includes several refinements of the simple “rights” expression of Elements. First, an “accessRights” field has been added. Although this clearly seems related to issues surrounding sharing data, sharing agreements do not aim to restrict access to content; rather, they are related to rules for appropriate ways of using content once it has been accessed. A “license” field has also been added as a subproperty of the “rights” term. This provides a place in the metadata that is more explicitly intended for storing a sharing agreement than the “rights” element of the original DC. It is, again, a free-text field, so expression of a sharing agreement as SPDX JSON is straightforward:

```
<?xml version="1.0" encoding="utf-8"?>
<metadata
xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance
xsi:noNamespaceSchemaLocation="http://purl.org/dc/terms/">
  [...]
  <rights>
    [...]
    <license>
        {
        "isDeprecatedLicenseId": false,
        "isFsfLibre": true,
        "licenseText": "Creative Commons Legal Code\n\nCC0 1.0 [...]",
        "standardLicenseTemplate": "[...]",
        "name": "Creative Commons Zero v1.0 Universal",
        "licenseId": "CC0-1.0",
        "seeAlso": [
          "https://creativecommons.org/publicdomain/zero/1.0/legalcode"
        ],
        "isOsiApproved": false
      }
    </license>
  </rights>
</metadata>
```

Figure 4.2 Sharing agreement expression in Dublin Core Terms

### *Sharing Agreements in FGDC CSDGM*

The Federal Geographic Data Committee (FGDC) established their Content Standard for Digital Geographic Metadata (CSDGM) in 1994, and its use was mandated by executive order for US federal agencies producing geospatial data (Executive Order 12906, 1994). Today, these agencies have been directed to transition to the use of ISO standard metadata (OMB 1998). Despite this transition, it is likely that significant catalogs of FGDC metadata will persist into the future, as translating or replacing existing metadata tends to be an activity of relatively high cost and low priority. The FGDC schema does not include elements specifically intended to contain information about modern sharing agreements. However, retrofitting existing FGDC metadata to express standard sharing agreements can be done within the constraints of the FGDC document definition and without making extensive changes to the document. Several elements of CSDGM are candidate locations for sharing agreement information:

§1.7, Access Constraints: restrictions and legal prerequisites for accessing the data set

§1.8, Use Constraints: restrictions and legal prerequisites for using the data set after access is granted

§6.3, Distribution Liability: statement of the liability assumed by the distributor

(FGDC, 1998)

Again, a sharing agreement describes activity taking place after accessing data, so Use Constraints is the more appropriate element for expressing a sharing agreement. Note that Distribution Liability is an element explicitly related to the distributor of data, who may or may not be the creator. A dataset may have multiple distributors, distributors can change over time, and each distributor may choose their own Distribution Liability statement, so this element is not an appropriate location to store sharing agreement information that is expected to endure as distribution

```
<?xml version="1.0" encoding="utf-8"?>
<metadata
xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance
xsi:noNamespaceSchemaLocation="https://www.fgdc.gov/schemas/metadata/fgd
c-std-001-1998.xsd">
  <idinfo>
    [...]
    <useconst>
      {
      "isDeprecatedLicenseId": false,
      "isFsfLibre": true,
      "licenseText": "Creative Commons Legal Code\n\nCC0 1.0 [...]",
      "standardLicenseTemplate": "[...]",
      "name": "Creative Commons Zero v1.0 Universal",
      "licenseId": "CC0-1.0",
      "seeAlso": [
        "https://creativecommons.org/publicdomain/zero/1.0/legalcode"
      ],
      "isOsiApproved": false
      }
    </useconst>
    [...]
  </idinfo>
  [...]
</metadata>
```

Figure 4.3 Sharing agreement expression for data in FGDC CSDGM

of the dataset changes. The dataset creator may express their own liability limitations within the sharing agreement, if needed.

Use Constraints is a free text field, meaning it can hold any characters that are permitted within an XML document. Above is an example of a minimal FGDC template containing a Use Constraints (useconst) section populated with the SPDX JSON expression of the CC0-1.0 waiver. The “licenseText” and “standardLicenseTemplate” elements of the sharing agreement have been omitted here for brevity using bracketed ellipses, but in practice may be included as written. (Other required portions of the metadata not relevant to this discussion have also been abbreviated here using bracketed ellipses.) The inclusion of this sharing agreement does not affect the validity of the metadata according to its XSD schema, and therefore a valid FGDC metadata remains valid when this content is included. A valid (but still minimal) FGDC metadata record example is included in the supplementary materials to this paper.

```
<?xml version="1.0" encoding="utf-8"?>
<metadata
xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance
xsi:noNamespaceSchemaLocation="https://www.fgdc.gov/schemas/metadata/fgd
c-std-001-1998.xsd">
[...]
  <metainfo>
    [...]
    <metuc>
      {
      "isDeprecatedLicenseId": false,
      "isFsfLibre": true,
      "licenseText": "Creative Commons Legal Code\n\nCC0 1.0 [...]",
      "standardLicenseTemplate": "[...]",
      "name": "Creative Commons Zero v1.0 Universal",
      "licenseId": "CC0-1.0",
      "seeAlso": [
        "https://creativecommons.org/publicdomain/zero/1.0/legalcode"
      ],
      "isOsiApproved": false
      }
    </ metuc >
    [...]
  </metainfo>
  [...]
</metadata>
```

Figure 4.4 Sharing agreement expression for metadata in FGDC CSDGM

The SPDX JSON representation of the CC0-1.0 waiver provides the full name and text of its CC origin document, also including a hyperlink to the full waiver on the CC web site. These elements provide ample human-readable access to the sharing agreement. Additionally, the “licenseId” element provides a unique identifier that serves as a key to the SPDX License List, allowing a machine process to unambiguously identify the specific sharing agreement in effect for the related dataset.

The FGDC schema also provides a location for sharing information referring to the metadata file, itself. Again, both Access Constraints and Use Constraints sections are available, but for the purpose of storing sharing agreements, the Use Constraints element is the more appropriate. Again, above see an XML snippet demonstrating the use of the SPDX JSON representation of the CC0-1.0 waiver to apply a sharing agreement to an FGDC metadata record.

### *Sharing Agreements in EML 2.2.0*

The Ecological Markup Language (EML) standard was developed at the National Center for Ecological Analysis & Synthesis (NCEAS) based upon the needs of the Ecological Society of America (ESA) and a foundational paper written by Michener et al. (Michener et al., 1997; Jones et al., 2019).

Version 2.2.0 of EML implements a section specifically intended to identify a sharing agreement that covers the data described by a metadata record (Jones et al., 2019). Within the “licensed” field of an EML metadata are elements for a “licenseName,”, “url”, and “identifier”. These fields align precisely with the Full Name, URL, and Identifier of specific licenses from the SPDX license list. Though the fields allow free text entry, they are explicitly intended for text referring to the SPDX license list. This native support of SPDX as a sharing agreement repository makes unambiguous the expression of sharing agreements in EML 2.2.0. One limitation of the EML “licensed” field is that it is intended to represent a sharing agreement covering both the metadata and the data for a particular dataset, limiting the ability of creators to express different sharing agreements for these different parts of a dataset.

```
<eml:eml> [...]
   <dataset> [...]
      <licensed>
         <licenseName>Apache License 2.0</licenseName>
         <url>https://spdx.org/licenses/Apache-2.0.html</url>
         <identifier>Apache-2.0</identifier>
      </licensed>
   </dataset>
</eml:eml>
```

Figure 4.5 Sharing agreement expression in EML 2.2.0

### *Sharing Agreements in EML Prior to Version 2.2.0*

The “licensed” element of EML was introduced with version 2.2, so earlier versions of EML do not provide for the expression of sharing agreements in that way. There are two ways of expressing access restrictions in earlier versions of EML, the “access” element and the “additionalMetadata” element. The “access” element, introduced in EML 2.1.0, is specifically used to “determine the level of access to a resource for the defined users and groups” and is intended for defining specific access controls within organizations in which potential users and groups are known. In practice, rules listed in the “access” element of EML may interact or conflict with more general sharing agreements, and care should be taken to harmonize rules expressed in multiple places in a metadata document. The “additionalMetadata” element is another place where sharing agreements can be expressed, and although the element is not specifically designed for this purpose, the EML project states that access rules have historically been stored there (Jones et al., 2019).

The additionalMetadata element contains a “describes” field that expresses the part or parts of the dataset to which the additional metadata applies, and a “metadata” element that contains the additional metadata text. The “metadata” field is a free-text field, and “allows EML to be extensible in that any XML-based metadata can be included in this element” (Jones et al., 2019). The “describes” field allows for a granular application of licenses to various components of a dataset described by EML. Using this method, it would be possible to list different sharing agreements (or none at all) for components individually identified within an EML metadata.

```
<eml:eml> [...]
   <dataset> [...]
      <additionalMetadata>
         <describes>[...]</describes>
            <metadata>
                <licenseName>Apache License 2.0</licenseName>
                <url>https://spdx.org/licenses/Apache-2.0.html</url>
                <identifier>Apache-2.0</identifier>
            </metadata>
      </ additionalMetadata >
   </dataset>
</eml:eml>
```

Figure 4.6 Sharing agreement expression in EML prior to 2.2.0

### *Sharing Agreements in ISO 19115-2*

The ISO 19115-2 metadata standard includes specific elements for expressing access restrictions for datasets described by the metadata. Within the “MD_Identification” section there is a “resourceConstraints” element that contains “MD_LegalConstraints,” a construct for listing and describing “accessConstraints,” “useConstraints,” and “otherConstraints.” The “accessConstraints” element, as in FGDC metadata, is used to describe restrictions to accessing the dataset, and is therefore not the most appropriate element for expressing a sharing agreement. The content of the “useConstraints” element is restricted to a specific code list, of which “license” is a member. The specification does not provide a field, free-text or otherwise, for describing the specific license unless the “otherRestrictions” code is chosen, but the element allows for multiple constraint codes to be chosen. Therefore, a complete expression for a sharing agreement would be to list the “license” and “otherRestrictions” codes for “useConstraints.” Provided that one or both of “accessConstraints” and “useConstraints” is populated with the “otherRestrictions” value of the code list, the “otherConstraints” field, a free-text field, may be populated with text describing the constraints more fully. The “otherConstraints” field can be populated with SPDX JSON to meet the needs of human- and machine-readability.

The ISO 19115-2 standard also supports applying the constraints elements independently to the metadata record, itself, within the “MD_Metadata” hierarchy of the record. Though the location of this section within the metadata record is different, the details of expression are the same, so the example provided below is for a sharing agreement applied to a dataset described by the metadata, rather than the metadata itself. In the full version of the example metadata included in the supplemental materials, sharing agreements are explicitly applied to both the dataset and the metadata.

In both the data and metadata constraints sections, the “otherConstraints” field is a single shared between access constraints and use constraints. That is to say, “otherConstraints” may contain information related to both access and use. Because there is no syntactical separation between “otherConstraints” text for access and use, there is the possibility for ambiguity, particularly in machine reading, between constraint text meant to apply to access and constraint text meant to apply to use. This ambiguity is resolved by a minor reorganization of elements in the ISO 19115-3 metadata schema.

```
<gmi:MI_Metadata> [...]
<gmd:identificationInfo> [...]
<gmd:MD_DataIdentification> [...]
<gmd:resourceConstraints>
<gmd:MD_LegalConstraints>
   <gmd:useConstraints>
      <gmd:MD_RestrictionCode
            codeList="http://www.ngdc.noaa.gov/metadata/published/
                     xsd/schema/resources/Codelist/
                     gmxCodelists.xml#MD_RestrictionCode"
            codeListValue="otherRestrictions">
                  otherRestrictions
      </gmd:MD_RestrictionCode>
   </gmd:useConstraints>
   <gmd:useConstraints>
      <gmd:MD_RestrictionCode
            codeList="http://www.ngdc.noaa.gov/metadata/published/
                     xsd/schema/resources/Codelist/
                     gmxCodelists.xml#MD_RestrictionCode"
            codeListValue="license">
                  license
      </gmd:MD_RestrictionCode>
   </gmd:useConstraints>
   <gmd:otherConstraints>
      <gco:CharacterString>
            {
            "isDeprecatedLicenseId": false,
            "isFsfLibre": true,
            "licenseText": "Creative Commons Legal Code\n\nCC0 1.0 [...]",
            "standardLicenseTemplate": "[...]",
            "name": "Creative Commons Zero v1.0 Universal",
            "licenseId": "CC0-1.0",
            "seeAlso": [
              "https://creativecommons.org/publicdomain/zero/1.0/legalcode"
            ],
            "isOsiApproved": false
            }
      </gco:CharacterString>
      </gmd:otherConstraints>
   </gmd:MD_LegalConstraints>
</gmd:resourceConstraints>
</gmd:MD_DataIdentification>
</gmd:identificationInfo> [...]
</gmi:MI_Metadata>
```

Figure 4.7 Sharing agreement expression in ISO 19115-2

### *Sharing Agreements in ISO 19115-3*

The expression of sharing agreements in the ISO 19115-3 standard is much like the ISO 19115-2 standard. Although the organization and namespaces of the metadata record are different, the basic

```
<gmi:MI_Metadata> [...]
<mdb:identificationInfo> [...]
<mri:MD_DataIdentification> [...]
<mri:resourceConstraints>
   <mco:MD_LegalConstraints>
       <mri:useConstraints>
             <mco:MD_RestrictionCode
             codeList="http://www.ngdc.noaa.gov/metadata/published/
                       xsd/schema/resources/Codelist/
                       gmxCodelists.xml#MD_RestrictionCode"
             codeListValue="otherRestrictions">
                   otherRestrictions
             </mco:MD_RestrictionCode>
       </mri:useConstraints>
       <mri:useConstraints>
             <mco:MD_RestrictionCode
             codeList="http://www.ngdc.noaa.gov/metadata/published/
                       xsd/schema/resources/Codelist/
                       gmxCodelists.xml#MD_RestrictionCode"
             codeListValue="license">
                   license
             </mco:MD_RestrictionCode>
       </mri:useConstraints>
       <mco:otherConstraints>
       <gco:CharacterString>
             {
             "isDeprecatedLicenseId": false,
             "isFsfLibre": true,
             "licenseText": "Creative Commons Legal Code\n\nCC0 1.0 [...]",
             "standardLicenseTemplate": "[...]",
             "name": "Creative Commons Zero v1.0 Universal",
             "licenseId": "CC0-1.0",
             "seeAlso": [
               "https://creativecommons.org/publicdomain/zero/1.0/legalcode"
             ],
             "isOsiApproved": false
             }
       </gco:CharacterString>
       </mco:otherConstraints>
   </mco:MD_LegalConstraints>
</mri:resourceConstraints>
</mri:MD_DataIdentification>
</mdb:identificationInfo> [...]
</gmi:MI_Metadata>
```

Figure 4.8 Sharing agreement expression in ISO 19115-3

module for constraint information is substantially similar. One notable change in ISO 1911-3 is that an “otherConstraints” element may be syntactically linked specifically with a “useConstraints” element in order to make explicit the link between the existence of a sharing agreement and the specific agreement in place.

## Future Work

Given the foundational role of copyright law on licenses, it would be valuable to perform a survey of existing data repositories to determine the prevalence of licenses and other sharing agreements applied to shared data as well as the country in which data were produced. Even though licenses may be adequate to cover data created in countries other than the US, questions remain about how those licenses might apply across national boundaries. In addition, a fresh look at sharing agreements from a legal perspective may be warranted, as statute and case law developed since the publication of the law review articles cited here may provide new insights into the legal status of data sharing agreements.

## Conclusion

Granting permission for people and algorithms to re-use data is critical to the future of open science. The best way to grant this permission is through the use of explicit sharing agreements. There are, at a minimum, two basic components to consider sharing agreements for when sharing data: a sharing agreement for the data themselves, and one for the metadata describing the data. Based upon copyright law limitations, data in the US likely do not qualify for coverage under the commonly used open source licenses, and must be released under a waiver of copyright rights. Metadata, as a “creative” work, could be shared under a license.

When choosing a metadata standard to describe data that are to be shared, care should be taken to select a standard that has the capacity to express sharing agreements for both the data and the metadata record, itself. When choosing a sharing agreement, commonly available agreements are a better choice than attempting to create new agreements; the creators of the existing agreements have spent years developing these agreements and some have the benefit of having been tested in court.

When sharing data using legacy metadata standards, we have shown methods for updating legacy metadata to include sharing agreements that are both machine- and human-readable. When sharing other materials alongside data products, such as software code, common open source licenses should be used as sharing agreements.

With a careful approach to metadata, license, and waiver selection, science data can be made available for re-use in unambiguous, explicit, and comprehensive terms.

## Acknowledgements

This material is based upon work that is supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture, under award number 2011-68002-30191; and by the ESIP Federation.

## References

Abelson, H., Adida, B., Linksvayer, M., & Yergler, N. (2012). ccREL: The Creative Commons Rights Expression Language. In M. Dulong de Rosnay & J. C. De Martin (Eds.), The digital public domain: Foundations for an open culture (pp. 149-187). Cambridge: Open Book Publishers CIC Ltd. doi:10.11647/OBP.0019

Bedini, I., Farazi, F., Leoni, D., Pane, J., Tankoyeu, I., & Leucci, S. (2014). Open government data: Fostering innovation. eJournal of eDemocracy and Open Government, 6(1), 69-79. Retrieved from http://www.jedem.org/index.php/jedem/article/view/329/271

Broussard, S. L. (2007). The copyleft movement: creative commons licensing. Communication Research Trends, 26(3), 3-17. Retrieved from http://cscc.scu.edu/trends/v26/v26_n3.pdf

Contreras, J. L., & Reichman, J. H. (2015). Sharing by design: Data and decentralized commons: Overcoming legal and policy obstacles. Science, 350(6266), 1312-1314. doi:10.1126/science.aaa7485

Copyright Law of the United States, 17 U.S.C. §105 (2016). Retrieved from https://www.copyright.gov/title17/92chap1.html#105

Gomulkiewicz, R. W. (2011). Enforcement of open source software licenses: the MDY Trio's inconvenient complications. Yale Journal of Law & Technology, 14(1), 106-137.

González, A. G. (2005). Open science: open source licenses in scientific research. North Carolina Journal of Law & Technology, 7(2), 321-366.

Guibault, L. (2013). Licensing research data under open access conditions under European law. In D. Beldiman (Ed.), Access to information and knowledge: 21st century challenges in intellectual property and knowledge governance (pp. 63-92). (Elgar intellectual property and global development). Cheltenham: Edward Elgar. doi:10.4337/9781783470488.00009

Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1(1), 1-136. Morgan & Claypool. doi:10.2200/S00334ED1V01Y201102WBE001

Hey, T., Tansley, S., Tole, K. (2009). The fourth paradigm: Data-intensive scientific discovery. Microsoft Cooperation. Available: http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf. Accessed 2019 Sep 30.

Jones, M. B., O’Brien, M., Mecum, B., Boettiger, C., Schildhauer, M., Maier, M., Whiteaker, T., Earl, S., Chong, S. 2019. Ecological Metadata Language version 2.2.0. KNB Data Repository. doi: 10.5063/F11834T2

Katz, Z. (2006). Pitfalls of open licensing: An analysis of Creative Commons licensing. Idea, 46(3), 391–413.

Khayyat, M., & Bannister, F. (2014). Open Data Licensing: More than meets the eye. Information Polity, 20(4), 231-252. doi:10.3233/IP-150357

Korn, N., & Oppenheim, C. (2011). Licensing open data: a practical guide. Higher Education Funding Council for England. Retrieved from http://discovery.ac.uk/files/pdf/Licensing_Open_Data_A_Practical_Guide.pdf

Lawrence, B., Jones, C., Matthews, B., Pepler, S., & Callaghan, S. (2011). Citation and peer review of data: Moving towards formal data publication. International Journal of Digital Curation, 6(2), 4-37. doi:10.2218/ijdc.v6i2.205

Lee, C. A., Allard, S., McGovern, N., & Bishop, A. (2016). Open data meets digital curation: An investigation of practices and needs. International Journal of Digital Curation, 11(2), 115-125. doi:10.2218/ijdc.v11i2.403

Leucci, S. (2014). Preliminary notes on open data licensing. Journal of Open Access to Law, 2(1). Retrieved from https://ojs.law.cornell.edu/index.php/joal/article/view/30/39

Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B., & Stafford, S. G. (1997). Nongeospatial metadata for the ecological sciences. Ecological Applications, 7(1), 330-342. doi: 10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2

Mockus, M., & Palmirani, M. (2015). Open government data licensing framework. In A. Kõ, & E. Francesconi (Eds.), 4th International conference on electronic government and the information systems perspective (pp. 287-301). Cham: Springer. doi: 10.1007/978-3-319-22389-6_21

Molloy, J. C. (2011). The open knowledge foundation: Open data means better science. PLoS Biology, 9(12), e1001195. doi:10.1371/journal.pbio.1001195

Muir, A. (2003). Copyright and licensing issues for digital preservation and possible solutions. In S. M. de Souza Costa, J. A. Carvalho, A. A. Baptista, & A. C. Santos Moreira (Eds.), Proceedings of the 7th ICCC/IFIP International Conference on Electronic Publishing, (pp. 89-94). Retrieved from https://elpub.architexturez.net/system/files/pdf/0315.content.pdf

Murray-Rust, P., Neylon, C., Pollock, R., & Wilbanks, J. (2010). Panton Principles, principles for open data in science. Retrieved from http://pantonprinciples.org

Odence, P., Lamons, S., & Lovejoy, J. (2013). Advancing the Software Package Data Exchange: An Update on SPDX. International Free and Open Source Software Law Review, 5(2), 145-152. doi: 10.5033/ifosslr.v5i2.89

Qin, J., & Li, K. (2013). How portable are the metadata standards for scientific data? A proposal for a metadata infrastructure. In International Conference on Dublin Core and Metadata Applications (pp. 25-34). Retrieved from http://dcpapers.dublincore.org/pubs/article/view/3670/1893

Sundara Rajan, M. T. (2011). Moral Rights: Principles, Practice and New Technology. Oxford University Press.

Silvello, G. (2018). Theory and practice of data citation. Journal of the Association for Information Science and Technology. doi: 10.1002/asi.23917

Software License Data Exchange. (2022, March 01). SPDX IDs. https://spdx.dev/ids/

Stodden, V. (2009). The legal framework for reproducible scientific research: Licensing and copyright. Computing in Science & Engineering, 11(1), 35-40. doi: 10.1109/MCSE.2009.19

Tenopir, C., Dalton, E.D., Allard. S., Frame, M., Pjesivac, I., Birch, B., et al. (2015) Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PLoS ONE 10(8):e0134826. doi: 10.1371/journal.pone.013482

Wilkinson, M. D., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018. doi: 10.1038/sdata.2016.18

Williams, A. J., Wilbanks, J., & Ekins, S. (2014). Why open drug discovery needs four simple rules for licensing data and models. In S. Moore (Ed.), Issues in open research data (pp. 77-88). London: Ubiquity Press. doi:10.5334/ban/

Yergeau, F., Bray, T., Paoli, J., Sperberg-McQueen, C. M., & Maler, E. (2008). Extensible Markup Language (XML) 1.0 (Fifth Edition). Retrieved from https://www.w3.org/TR/xml/

# Chapter 5: Conclusion

## Introduction

Our traditional concept of the practice of science, from research proposal to publication, has been a wildly successful endeavor, as evidenced by the proliferation of high-quality journals and articles. Modern communications and computational technology have come a long way in a short time, and high-speed networking and telecommunications have made it easier to collaborate and to share research data with colleagues across the world. These capabilities are a relatively recent development, and as with any new frontier, there has been a period of time in which it is difficult to know how to work efficiently, how to construct and work within a framework for interaction that helps to establish clear boundaries, how best to take advantage of new possibilities that have come with these advances in technology.

Big Data, ill-defined as it is, can be a good stand-in for many of the problems we face with modern research. The three Vs of Big Data, volume, variety, and velocity, all typify the explosion of opportunity afforded by technology. Data are available today in volumes that were practically impossible to manage only a few decades ago. Global data production is measurable in zettabytes (a zettabyte is a billion terabytes) per year. In terms of data variety, where telecommunications links once restricted users primarily to text transmissions, today data are transferred as text, images, audio, video, and every combination of these, constantly, every day. Of course, velocity may refer simply to our faster networks, but it can also refer to the speed with which data are entering the network in parallel due to the proliferation of people with access to the Internet. Not only must we deal with the data deluge in practical terms of where, how, and whether to store all these data, but we must also develop common technologies and understandings for shared access.

Technologies such as repositories help to guide and enforce the understandings that we must share to take advantage of Big Data. There is a clear ideology behind the OAIS model. The very concept of a common model for repository design implies, at least to an extent, common interactions, common policies, and common goals. There would be no need for open repository standards if we did not plan for our systems to interact. There would be no need for repositories without a recognition of the value of the data lifecycle—not just as it applies to a single project, but to the cyclical nature of it, to the potential for mining data products for value over and over again. In fact, the purpose of a standard is to be applied repeatedly, as is the OAIS repository model. The implementation of the OAIS model as a service-oriented architecture provides the opportunity not just to re-use the model itself, but also to re-use the modules contributing to the implementation of the model. As we enable the re-use of data in open science, we can also enable the re-use of tools that make these methods possible.

The opportunities arising from Big Data also present approaches to treating the  recent replicability crisis, although it has been with us in some form in the sciences forever. There are many reasons a result may be impossible to replicate: poorly or incompletely explained procedures, common statistical error, failure to account for multiple comparison, sample sizes that overpower statistical tests, and so on. Perhaps one of the simplest explanations for the replicability crisis was given by Feynman in a 1974 commencement address at CalTech: “The first principle is that you must not fool yourself — and you are the easiest person to fool.” Technologies such as open science principles can make it harder for researchers to fool ourselves, because they force us to do our research very much in the open, to subject our data and our methods to scrutiny that wasn’t practical without the Internet. Again, we impose ideology through technology. Whether or not the exact approach to open science demonstrated in chapter 3 becomes the preferred method of practicing open science, by implementing the concepts in any form, we advance the conversation and provide new ways to discuss the concept. The specific mode of implementation is not as important as the fundamental idea that science must be repeatable, and that open science provides repeatability through transparency.

Technologies like sharing agreements are also necessary. As the culture of data sharing matures, rules must also mature, shifting from the ad-hoc and unwritten to the standard, explicit, and clearly expressed. In all spaces in which people interact and compete, disputes will occur, and some will be processed within the legal system. Rules will be tested. The use of standard rules designed by legal experts for the purpose of expressing sharing agreements regarding data seems an approach likely to withstand testing. In the same ways that repositories and open science impose ideology, so do these sharing agreements, quite explicitly, in the terms they adopt. The data creators who choose the sharing agreements are also making an ideological statement. It is important that data creators are thoughtful about the sharing agreements they choose, in order to ensure that they are effective once data are released, and the added benefit of making sharing agreements machine-readable is that open science takes steps toward automation.

These chapters have addressed repository design; open science, repeatability, data re-use; the FAIR data principles and sharing agreements. Together, these concepts form facets of the shape of the solutions to modern challenges in the sciences related to Big Data. Each also contributes to establishing a common ideology of modern science, through obvious channels such as the text of sharing agreements and less obvious ways, like choices made in the definition of a repository API function. The tremendous challenges of the era of Big Data also bring great opportunities to improve upon ourselves and upon our disciplines and make science work better.

## Future Work

The natural evolution of the data repository is a system that provides access not only to open data that are interoperable and ready to use, but also software systems that enable complex analyses that can be remotely defined and performed on data in situ. As repositories grow in technical capability, and as data production matures to adhere to standards in data organization (like standards for recording observation and measurement) and metadata production, the additional structure of data collections begins to enable data interoperability. Our standards for repositories are already enabling federated activities—such as federated search systems that consume and index the content of multiple discrete repository systems. In principle, federated activities are not limited to search. As emerging methods of machine learning, or weak artificial intelligence, develop, we are creating analytical systems capable of finding solutions to complex problems that defy traditional algorithms. The interoperable nature of structured data will make combining repository data with machine learning an obvious next step. The systems needed to make these connections already exist, at least in nascent form.

The future of work in this field is to continue to grow and shape the technologies surrounding data collection and organization and the technologies of machine learning to architect data systems that are capable not simply of warehousing data, but of analyzing data and producing scientific outputs on request.