Practices, Challenges and Prospects of Big Data Curation: A Case Study in Geoscience

Suzhen Chen
University of Hawai'i at Mānoa

Bin Chen
University of Hawai'i at Mānoa

Abstract

Open and persistent access to past, present, and future scientific data is fundamental for transparent and reproducible data-driven research. The scientific community is now facing both challenges and opportunities caused by the growingly complex disciplinary data systems. Concerted efforts from domain experts, information professionals, and Internet technology experts are essential to ensure the accessibility and interoperability of the big data. Here we review current practices in building and managing big data within the context of large data infrastructure, using geoscience cyberinfrastructure such as Interdisciplinary Earth Data Alliance (IEDA) and EarthCube as a case study. Geoscience is a data-rich discipline with a rapid expansion of sophisticated and diverse digital data sets. Having started to embrace the digital age, the community have applied big data and data mining tools into the new type of research. We also identify current challenges, key elements, and prospects to construct a more robust and future-proof big data infrastructure for research and publication for the future, as well as the roles, qualifications, and opportunities for librarians/information professionals in the data era.

Received 06 May 2019 ~ *Accepted* 11 September 2019

Correspondence should be addressed to Suzhen Chen, Cataloging Department, University of Hawai'i at Mānoa Library, 2550 McCarthy Mall, Honolulu, Hawaii 96822. Email: suzhen@hawaii.edu

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

 doi:10.2218/ijdc.v14i1.669

Introduction

The past two decades have been witnessed by exponential growth of large experimental, observational and simulation datasets, often beyond the petabyte and exabyte levels, with increasing complexity and diversity in domain repository across many scientific disciplines. This trend has unprecedentedly revolutionized the way scientific discovery are made in the new era, but has also posed new challenges in acquiring, organizing, and processing the huge and rapidly growing data systems (Guo, 2015; 2017). Simply put, big data refers to a large volume of data or substantial collections of data (Davenport, 2014). The source of data can be in a variety of ways, such as bibliographic data, unstructured data and research data, each of which may have different data structure in a variety of collections or datasets as well as its own complexity, volume and quality (Haynes, 2018).

Accessing past, present, and future scientific data is of paramount importance to make scientific research transparent and reproducible, so that the products of past and current research can be re-used to empower future science, and thus benefit the society (Lehnert and Hsu, 2015). Big data is not only the collective sum of small data but can help achieve far greater contributions to research than all of its parts; more importantly, it may generate new findings, applications, and solutions that cannot be possible from its constituent subsets of the data (Haynes, 2018). The data-centered activities culminate in total value and can be described by a data value chain through data discovery, integration, and exploitation processes, which also illustrates from raw data to decision making the interdependent relationship between stages (Miller and Mork, 2013). It has been widely recognized that the value of open and persistent data grows as they become discoverable, citable, re-usable, integrated, and linked with other data (Lehnert and Hsu, 2015).

Scientists and the scientific community as a whole must prepare themselves to welcome the new era of data-intensive scientific research and discovery. Geosciences, in particular, is a traditionally descriptive and field centric discipline. Providing precise descriptive metadata of the field and experimental data is essential for transition into modern digital scholarship practices. Mechanisms, infrastructure, and incentives to transition into modern digital scholarship practices are trending in the data-intensive geoscience field (Gil et al., 2016). Figure 1 demonstrates the increasing interest in the usage of big data in science and the geoscience field, based on the data available in Google Trends from 2004 to 2018. It shows the significant growth in the interest in big data in science and geoscience since 2012.

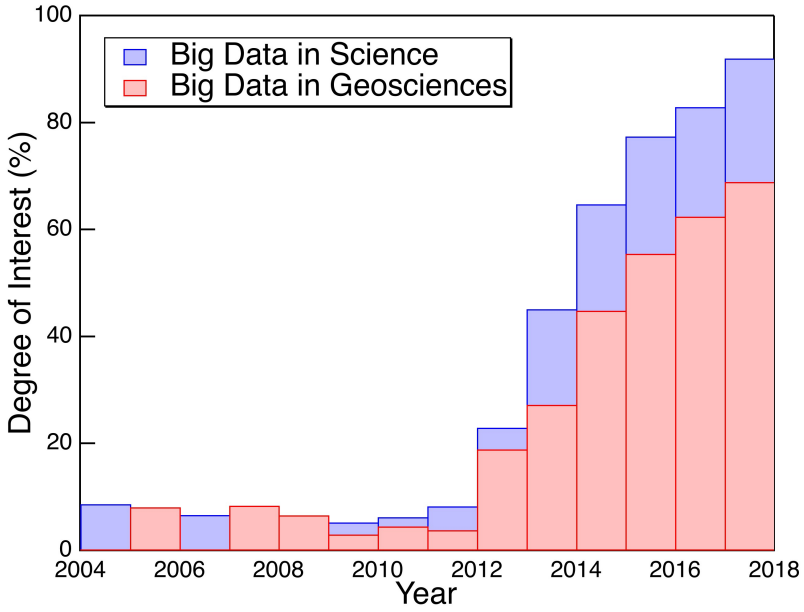


Figure 1. Degree of interest on big data in science and geosciences (based on Google Trends and acquired in January 2019).

Challenges and prospects of big data are increasingly great, as the volume and value exponentially cumulate over time. With more data submitted to a domain repository where they are properly curated and maintained by librarians or other information professionals, the whole scientific community will benefit from the summed benefit of big data. In an academic setting, there is a growth in the needs of big data in the research fields and an increasing need to have a better understanding of big data for research purpose, in the perspectives of data accessibility, consistency, and interoperability. Despite the importance of big data for data-intensive scientific discovery, the theory, methodology, and models, as well as roles of involved experts including librarians/information professionals, IT experts and end users/domain experts, have not been reviewed in depth. In this paper, we will review the practices, application, and challenges of big data using geoscience as a case study. We will discuss the roles and opportunities of librarians/information professionals in enhancing the value of the big data.

Practices of Scientific Big Data: A Case Study in Geosciences

Geoscience is a discipline that spans from the core of the Earth to the top of the mountains, and to the stratosphere, and to the interplanetary space. Geoscientists have long embraced the benefits from larger, more diverse datasets for decades to solve sophisticated and challenging geoscience problems and taken advantages of the advances in measurement/data acquisition technology and the improvements in computing power. As an example for the case study, Interdisciplinary Earth Data Alliance (IEDA) is a data facility funded by the USA National Science Foundation (NSF) to provide data repositories, data syntheses, metadata catalog, and data visualization and analysis tools for Ocean, Earth, and Polar Sciences (Lehnert et al., 2018). The data

hosted in IEDA has been shared and extensively used by geoscience researchers from different disciplines in geosciences. The NSF-supported EarthCube[1] initiative, on the other hand, focuses on building community-driven cyberinfrastructure for managing, sharing, and exploring geoscience data and information to enable data-driven scientific discovery (Black et al., 2014). To enhance the scientific value of geoscience big data, the following practices have been adopted by the community to ensure the success of the cyberinfrastructure of big data.

Data Curation by Domain Experts, Information Experts, and IT Experts

For decades, it has become a strong trend that large and diverse databases demand concerted effort from geoscience application scientists and information science experts (Ebert-Uphoff et al., 2017). The collaboration between librarians/information professionals and geoscientists/researchers becomes vital to develop and enhance metadata profiles for geoscience data. There is a closer and richer collaboration between geoscientists and information professionals in the data curation, data management, and data sharing. Larger and more diverse datasets need sophisticated mathematical and computer science expertise (Ebert-Uphoff et al., 2017).

Figure 2 is an adapted version of the illustration of the collaboration between knowledge creation, organization, and maintenance, which shows the partnership between "domain experts, information experts, and information technology experts" (Calhoun, 2007). In geosciences, domain experts are typically geoscientists who acquire, create, and use the data. Information experts often refer to librarians, records managers, archivists, data scientist, or other information professionals who provide technical services or data/records management to facilitate learning and research, by selecting, acquiring, organizing, and preserving information in systems and structures, as well as enhancing the accessibility and quality of information. Information technology (IT) experts refer to the specialized professionals who build and maintain cyberinfrastructures for the geoscientific data repositories.

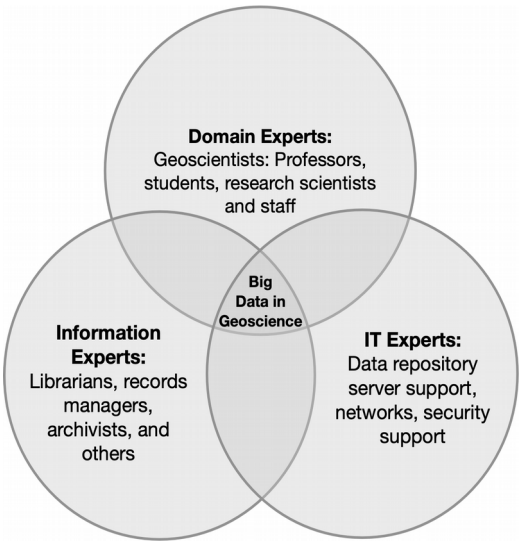


Figure 2. Geosciences big data pyramid, modified after Calhoun (2007).

1 EarthCube: http://www.earthcube.org

Metadata, Controlled Vocabulary, and Thesaurus for Geosciences Data

Metadata, defined as "data about data", plays an increasingly important role in information organization and management (Riley and National Information Standards Organization (U.S.), 2017). Metadata are structured or encoded data that describe the characteristics of the entities to assist in the identification, discovery, assessment, and management of the entities (American Library Association (ALA) Committee on Cataloguing: Description and Access, 2000). Haynes identified five purposes of metadata, which include resource identification and description, information retrieval, information resources management, information rights management, learning, research and commerce support, and information governance (Haynes, 2018). It is essential to ensure the quality of metadata to aid in data reuse and big data organization. Poor metadata will hinder the interoperability and long-term sustainability of data (Sweet and Moulaison, 2013). This is echoed by metadata researchers such as Zeng and Qin, who stated that the quality of metadata is essential for the usefulness and usability of a collection of resources and discovery system to ensure satisfaction in the information searching process (Zeng and Qin, 2016). Metadata indicators, such as completeness, correctness and consistency, are closely linked to metadata quality and data interoperability (Sweet and Moulaison, 2013).

Metadata standards are essential for the discoverability, sharing, and reuse of geoscience datasets. However, there exists a lack of metadata standards developed for experimental and observational datasets in the geoscience field. Specifically, metadata for laboratory data are challenging to analyse, construct, and access. Some of the experimental data are open-ended and it is challenging to develop metadata fields to apply to every type of experiment (Lehnert et al., 2015). Geoscientists have recognized the lack of norms or standards of data reporting in publications. Of particular note, the geochemistry community has worked together to define a data-reporting norm or standards for U-series geochronology data in literature, in order to standardize the current data-reporting practices (Dutton et al., 2017). That is, community-defined minimum data and metadata can serve as a community norm/standard to aid in later use, whereas suggested additional information would allow subsequent reanalysis.

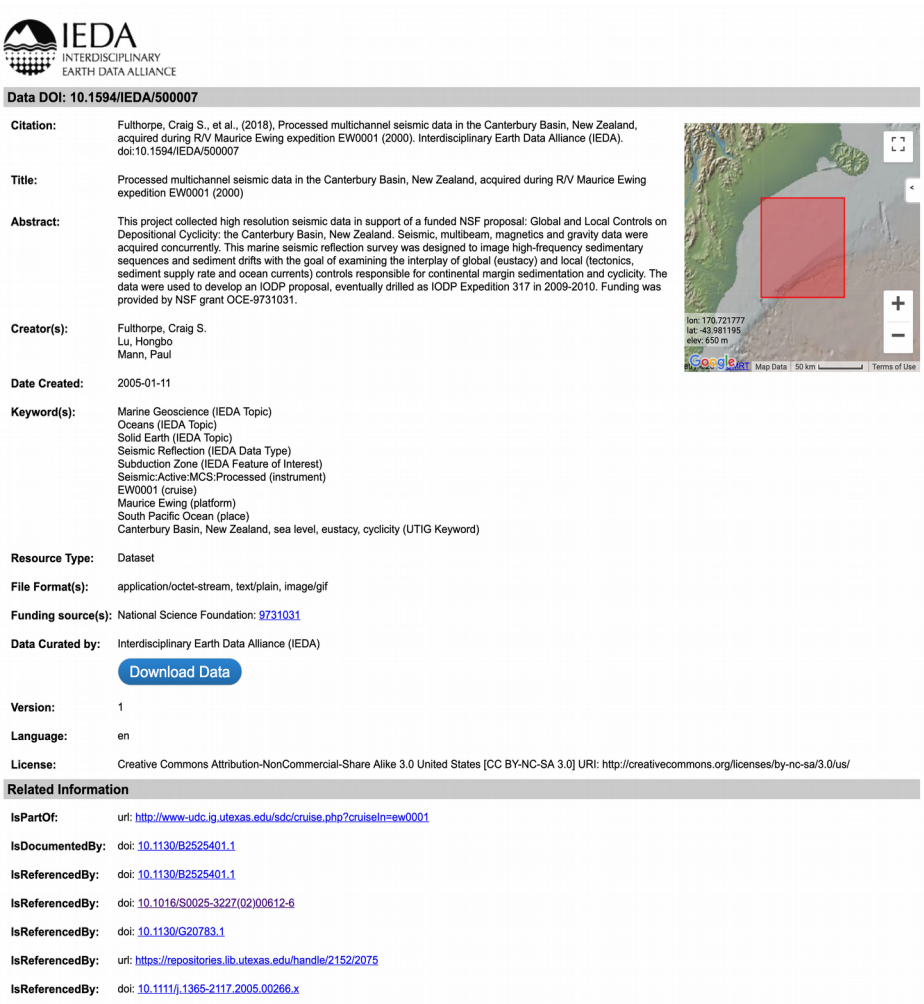
Metadata play an essential role in big data management and facilitate data interoperability. Controlled vocabulary and data consistency are crucial to facilitate the geoscience researchers to use the data. Using IEDA as an example, it hosts data repositories from different disciplines such as Marine Geoscience Data System, System for Earth Sample Registration, Petrological Database, etc. Each repository contains vocabularies, authority files, and hierarchies. To ensure the databases, software tools, and workflows follow community-based standards and adopt the best practices for samples metadata, classification, identification, and registration, the IEDA community has developed a thesaurus: a controlled vocabulary to combine separate controlled vocabularies from different systems to a single master-controlled vocabulary (Ji et al., 2014; Lehnert and Hsu, 2015). The IEDA thesaurus contains 18 top facets, which include equipment, geographic gazetteer, geologic ages, geologic units, materials, etc. (Ji et al., 2014). The thesaurus of IEDA is organized along with the "ANSI/NISO Z39.19-2005 Guidelines" for the Construction, Format, and Management of Monolingual Controlled Vocabularies, and is published using Simple Knowledge Organization System (SKOS) format. The IEDA thesaurus server provides classic web semantic features, such as SPARQL, RESTful web services, and unique URI based on open source technologies. In the long term, data needs to be secured and stored by institutions

and domain repositories. To allow for data interpretation, scientists need to join forces in defining community standards for the disciplinary data, which includes defining and recording appropriate metadata, such as experimental parameters and set-up (Lynch, 2008). On the other hand, ensuring the quality of the data is somewhat more important than the amount of the data and the integration of the data, which will also help the interoperability of the various databases and systems and integration with other data types for interdisciplinary research (Miller and Mork, 2013).

More effort should be made in constructing metadata, including descriptive metadata, structural and administrative metadata in geoscience for different databases. They should be employed consistently when geoscientists deploy the data. In order to optimize cyberinfrastructure capabilities for samples and sample-based data – especially integration with other data types for interdisciplinary research – databases, software tools and workflows need to follow community-based standards and best practices for sample metadata, classification, identification, and registration. For example, a new EarthChem database is under development using the ODM2 information model (ODM=Observation Data Model) for spatially discrete, feature-based Earth observations that integrate observations from in-situ sensors and environmental samples, aligned with OGC's Observation and Measurements model (Horsburgh et al., 2014).

Persistent Identifiers for Research Data

The use of persistent identifiers for research data has been recognized as a paramount issue for data publication and citation. A resource, including scientific data, is often accessed on the Internet by its Universal Resource Locator (URL), the so-called "web address". However, URL often changes over time due to the reorganization of web servers, so that the link to the data resources may get lost and persistent access to the resource has become a serious problem for the valuable resources scattered around the Internet. A Digital Object Identifiers (DOI) is a URN (Uniform Resource Name), consisting of a compact string for providing a unique, persistent, and actionable identifier of a digital object (DeRisi et al., 2003). It typically consists of a publisher ID (prefix) and an item ID (suffix), separated by a forward slash (/). The DOI system is governed by the International DOI Foundation (IDF). Simply put, the DOI is used for redirection (called "resolution") from a persistent identifier to a URL. The advantage of DOI over URL or persistent URL (PURL) is their best archival guarantees. Even though the URL changes or an object moves, its DOI remains the same, so that the current location of the associated object can be easily updated in the international registry (DeRisi et al., 2003).



IEDA
INTERDISCIPLINARY EARTH DATA ALLIANCE

Data DOI: 10.1594/IEDA/500007

Citation:	Fulthorpe, Craig S., et al., (2018), Processed multichannel seismic data in the Canterbury Basin, New Zealand, acquired during R/V Maurice Ewing expedition EW0001 (2000). Interdisciplinary Earth Data Alliance (IEDA). doi:10.1594/IEDA/500007
Title:	Processed multichannel seismic data in the Canterbury Basin, New Zealand, acquired during R/V Maurice Ewing expedition EW0001 (2000)
Abstract:	This project collected high resolution seismic data in support of a funded NSF proposal: Global and Local Controls on Depositional Cyclicity: the Canterbury Basin, New Zealand. Seismic, multibeam, magnetics and gravity data were acquired concurrently. This marine seismic reflection survey was designed to image high-frequency sedimentary sequences and sediment drifts with the goal of examining the interplay of global (eustacy) and local (tectonics, sediment supply rate and ocean currents) controls responsible for continental margin sedimentation and cyclicity. The data were used to develop an IODP proposal, eventually drilled as IODP Expedition 317 in 2009-2010. Funding was provided by NSF grant OCE-9731031.
Creator(s):	Fulthorpe, Craig S. Lu, Hongbo Mann, Paul
Date Created:	2005-01-11
Keyword(s):	Marine Geoscience (IEDA Topic) Oceans (IEDA Topic) Solid Earth (IEDA Topic) Seismic Reflection (IEDA Data Type) Subduction Zone (IEDA Feature of Interest) Seismic:Active:MCS:Processed (instrument) EW0001 (cruise) Maurice Ewing (platform) South Pacific Ocean (place) Canterbury Basin, New Zealand, sea level, eustacy, cyclicity (UTIG Keyword)
Resource Type:	Dataset
File Format(s):	application/octet-stream, text/plain, image/gif
Funding source(s):	National Science Foundation: 9731031
Data Curated by:	Interdisciplinary Earth Data Alliance (IEDA)
	Download Data
Version:	1
Language:	en
License:	Creative Commons Attribution-NonCommercial-Share Alike 3.0 United States [CC BY-NC-SA 3.0] URI: http://creativecommons.org/licenses/by-nc-sa/3.0/us/

Related Information

IsPartOf:	url: http://www-udc.ig.utexas.edu/sdc/cruise.php?cruiseIn=ew0001
IsDocumentedBy:	doi: 10.1130/B2525401.1
IsReferencedBy:	doi: 10.1130/B2525401.1
IsReferencedBy:	doi: 10.1016/S0025-3227(02)00612-6
IsReferencedBy:	doi: 10.1130/G20783.1
IsReferencedBy:	url: https://repositories.lib.utexas.edu/handle/2152/2075
IsReferencedBy:	doi: 10.1111/j.1365-2117.2005.00266.x

Figure 3. An example of an IEDA data publication with a DOI, metadata of the data sets, license, as well as related information such as studies citing the data set.[2]

Scientific articles nowadays often have an associate DOI for unique identification. Since 2014, assigning DOIs to research data has become an integral part of data publication and citation (Klump et al., 2016). DOIs are only issued and maintained by authorized sites; DataCite[3], a non-profit organization, was founded to govern the system for the assigning of DOI for research data and develops supporting technology. IEDA offers a data publication service that registers geoscience data in the DOI system through the DataCite consortium, in order to make datasets accessible and citable as publications with attributions to the contributors as authors (Figure 3). Figure 3 shows an example of a dataset or data publication in IEDA. The data publication is citable with a DOI "10.1594/IEDA/500007" and detailed citation information including author and title of the datasets. The metadata of the data publication includes "Abstract", "Creators", "Date Created", "Keywords", "Resource Type", "File Format(s)", "Funding Source(s)", "Data Curated By", "Version", "Language", and "License". There is also a section on "Related Information", containing information on the project that this dataset belongs to, the document that describes the datasets, and papers/reports (with URL or DOI) that cites the datasets.

2 See http://get.iedadata.org/doi/500007 for the data publication in IEDA
3 DataCite: https://datacite.org/

Data Publication and Incentives for Contributing Data

Scholarly publication is, at present, a preferred way of publishing scientific data. When data volumes are small, they can be included in articles as data tables and/or electronic supplements (Lehnert and Hsu, 2015). Data published in this venue, however, are highly dispersed and lack critical metadata and compliance with data standards, making finding, accessing, mining, and reusing the data extremely difficult. Data publication has emerged as important part of scientific workflows, research communication, and scholarly values to make data open and accessible; it has been encouraged and promoted by governments, funding agencies, and academic institutions, professional societies, and publishers (Lehnert and Hsu, 2015).

In geoscience, one way to share data is to submit data by researchers to remote repositories that can be shared, preserved, discovered and re-used. The end users will need to spend significant time in organizing, describing, and uploading their data to the domain repository. How can this type of data publication and contribution be recognized and rewarded? One such strategic move is to give stakeholders incentives to share their data. The creators of the data and software need to properly be credited for the effort through assigning unique identifiers and citations.

Proper data citation is crucial for providing an incentive to encourage data publication, sharing, and reuse (Mooney and Newton, 2012). Incentives also include recognition of impactful informatics by peer committees and research-rating exercises (Nature, 2008). The incentives can also come from support from the universities and funding agencies for curation facilities, tools, and trainings (Nature, 2008). As an example, the EarthChem Library[4] is an open-access repository for geochemical datasets as part of IEDA. Its various data repositories provide long-term archiving and registration of data with DOIs, which can be cited by research articles. The number of citations to the data records hosted in databases such as "EarthChem", "PetDB", and "IGSN/SESAR" have steadily increased since their establishments (see Figure 4), enabling data-intensive research for the community.

4 EarthChem: http://www.earthchem.org

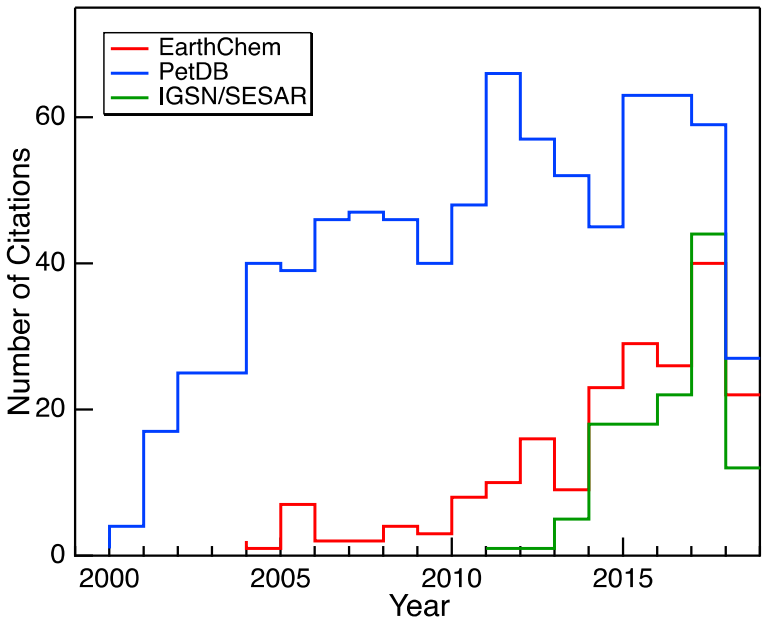


Figure 4. Number of citations from the years when the IEDA databases "EarthChem", "PetDB", and "IGSN/SESAR" were established.

Another type of incentive for researchers to contribute data to the repository is to offer awards to support such an effort. IEDA, for example, gives out mini-awards to individuals or teams to rescue long-tail data, which are small-volume and project-specific scientific data produced by individuals and small teams (Hsu, Lehnert, et al., 2015). Those long-tail data typically lack adequate metadata for sharing between teams and individuals. Therefore, the rescue effort supported by the mini-awards entails well-structured and adequate metadata for data reuse.

Challenges and Prospects of Big Data in Geoscience

Long-Term Digital Preservation and Big Data Curation

Long-term storage and curation of big data require the sustainability for securing public access over a long period of time. Currently, there are typically three types of data storage solutions: discipline-specific repositories, institutional repositories, and commercial cloud storage systems (Hsu, Martin, et al., 2015). One major challenge of long-term data storage and curation is the lack of a clear blueprint for funding support. It is hard for an individual researcher to fund long-term data management and curation for a regular research proposal. One solution could be a national or international discipline-specific data storage and curation infrastructure funded by research agencies. Specifically, IEDA is a good example of discipline-specific storage solution. Depending on the purpose and discipline scope, individual IEDA systems were developed independently and operated with disciplinary focus and expertise, some of which are

actively maintained data synthesis for advanced data mining and analysis (Carter-Orlando et al., 2017). In addition, IEDA offers data submission and access interfaces to streamline the submission process, as well as data visualization and analysis tools to improve the usability of its portfolio.

Researchers may also use the data infrastructure in their universities or institutions for long-term storage solution funded by overhead or general budgets. For instance, the University of California Digital Preservation Program aims to provide long-term curation and access to digital assets provided by university-affiliated agents (Abrams et al., 2009). The Program employs a deliberative multistage design process, from value, to strategy, to service, and to system, in order to allow the curation activities to be proactive rather than reactive. However, not every institution has such a digital library program for long-term data storage and curation. Alternatively, commercial cloud data storage service could be a cost-effective option for individual researchers, but long-term funding framework may be an issue (Hsu, Martin, et al., 2015).

Understandable Geoscience Context and the Geoscience Paper of the Future

Today, datasets are often published with an article as supplementary materials in many geoscience journals. However, the journals often do not require metadata entailing the connections between the data, software, and results for reproducible and transparent research. Geoscience data by nature has a rich background and requires necessary summary and descriptions of the context, ranging from the scientific motivations for collecting the data, the instruments for acquiring the data, the pre-processing analyses of the data, and the scientific results (Ebert-Uphoff et al., 2017). The challenges of building big data repository systems include the diversity in data and records from different research areas in geosciences. Formulating the guidelines for specific metadata for data sets collected in the field, experiments, and simulations can be challenging. Typically, efforts for standardized metadata documentation are made by the whole community for developing shareable linked datasets (Moulaison et al., 2012). Though scientific data is not easy to analyze, Hsu, Martin, et al. (2015) suggested using vocabularies established in the Earth science community, such as CSDMS standard names. Using controlled vocabulary can facilitate data sharing, which makes data more accessible, findable, and interoperable. There has been a realization of the importance of controlled vocabularies in the geoscience field. Efforts have been put to develop controlled vocabularies. Some of the commonly used thesauri for the geoscience field include GeoRef Thesaurus, Global Change Master Directory (GCMD), U.S. Geological Survey Library Classification System, and Semantic Web for Earth and Environmental Terminology (SWEET) ontologies (Ji et al., 2014).

Reproducibility and quality control of experimental, field, and observational data have become two crucial issues for data-intensive geoscience research (Hsu, Martin, et al., 2015). In geosciences, some time-sensitive measurements and observations cannot be repeated due to the time-sensitive nature of the data. Therefore, replicate tests and measurements could not be performed. Some measurements cannot be repeated due to the limited funding and lack of repeatable procedures and workflows, although replicate experiments may significantly help minimize the uncertainty associated with experimental methods and measurements protocols (Hsu, Martin, et al., 2015). In turn, quality control is hampered by the lack of reproducibility. Poor reproducibility and quality control of the geosciences data may limit researchers' ability to perform data-centered research (Hsu, Martin, et al., 2015).

Open and persistent access to past, present, and future scientific data is fundamental for transparent and reproducible data-driven research. The value of the data will be enhanced while being reused by others and thus amplify the research potential of shared data and software. Future publications of scientific papers will not only publish text and figures but also the computational workflow and the associated digital objects, such as data and software in the digital age. In order to improve scientific communication, give credit to scientific contributions, and promote the reproducible and transparent research, reproducibility and computational provenance will be key review criteria for future geoscience publications (Gil et al., 2016). Gil et al. (2016) emphasized the importance of open science (data and software), reproducibility, and modern digital scholarship for Geoscience Papers of the Future (GPF). The suggested best practices for GPF are reusable data, software, and computational provenance of results through publication in a public repository with necessary metadata, license, and unique and persistent identifier for citation (Figure 5).

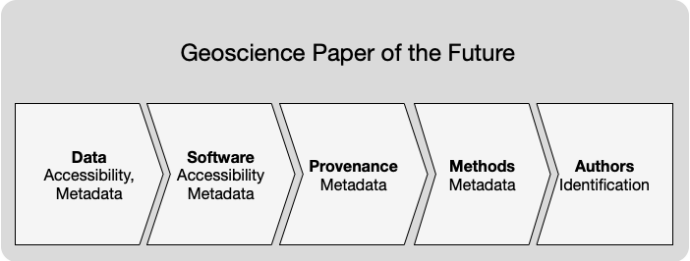


Figure 5. A Geoscience Paper of the Future (GPF) includes data, software, and computational provenance as expected in reproducible publications but also include desirable features in open science and digital scholarship. Modified after Gil et al. (2016).

Linked Data and Data Interoperability

Use of the proper descriptive standards for data records is important for ensuring the quality and interoperability of large datasets. Next-generation data-intensive research to understand the Earth as a system motivates the need for integrating and interlinking of vast and complex datasets from multiple domains (Yu and Liu, 2015). The essence of Linked Data is to allow data, tools, and models to connect with each other to form a "data network" and achieve data interoperability. The term normally used to define the set of features that data or metadata need to have in order to allow for this linking and combining of heterogeneous data is "data interoperability", which is a feature of datasets and of information services that give access to datasets, whereby data can easily be retrieved, processed, re-used, and re-packaged and re-operated by other systems (L'Abate et al., 2015).

Using the Linked Data approach is a paradigm shift for the geoscience data infrastructure and Web-scale data integration. Taking IEDA as an example, one of its major foci in the near future is to network internationally with other disciplinary data systems, such as geochemical databases operated by providers in Germany (GEOROC) and in Japan (GANSEKI) (Lehnert et al., 2018). The interoperability of the individual domain repositories in geoscience curated by various institutions or data alliances requires Linked Data for an enhanced integrated global data resource. The fundamental data structure of ontologies and datasets in the Semantic Web is the Resource

Description Framework (RDF)[5] which has a triple for "Subject, Predicate, Object" (Ma, 2017). Through the use of ontologies, another NSF-supported next-generation cyberinfrastructure for geosciences, EarthCube, has a building block project GeoLink, which focuses on improving data retrieval, reuse, and integration of participating geoscience data repositories.

Community Engagement

Even when all the infrastructures exist for sharing geoscience data, there still exist significant cultural and institutional challenge that may impede the open sharing of data (Hsu, Martin, et al., 2015). NSF requires that a "Data Management Plan" must be included as a supplementary document in every proposal submitted to the agency (National Science Foundation, 2019). Other agencies, such as NASA, also require data management plans. Further, NSF released a public access plan entitled "Today's Data, Tomorrow's Discoveries", promoting increased access to the results of funded research in 2015 (National Science Foundation, 2015). Data sharing pertains to not only data sets, but also research tools developed under grant-funded research (Diekema et al., 2014).

Information technology has advanced rapidly over the last decade, foreseeing extraordinary progress and disruptive change in data-intensive research and discovery in the new era. However, there exist several challenges that need to be overcome for the paradigm change in data science research. For geoscience, in particular, many field-based geoscientists do not have the time and skills to create metadata, document workflow, and submit the data into the online repositories (Hsu, Martin, et al., 2015). It may not be in the self-interest of a researcher to take the time to gain the skills and eventually share their data, as sharing data are not currently part of the evaluation and recognition of the work by a researcher (Diekema et al., 2014). Domain scientists who collect the data do not necessarily have the expertise or awareness for data management and curation. Although training materials are available in various locations, such as university libraries, funding agencies' website, etc., those resources are typically new, and scientists may not be aware of the existence of those materials and motivated to learn more. Therefore, the lack of incentives for researchers to contribute data sets, software, and provenance is a paramount problem for data publication.

The synergy between domain experts, information professionals, and IT experts are paramount in big data management and applications. Domain experts are typically on the side of the spectrum of contributing, discovering, and reusing the data, whereas information experts and IT experts are curators and facilitators in the process of big data management, respectively. How to evaluate the success of a big data cyberinfrastructure (CI)? Cutcher-Gershenfeld et al. (2016) reported a baseline assessment of engagement with the NSF EarthCube Initiative, an open CI effort for the geosciences, based on survey results of geoscientists and CI experts. Based on the survey data, they found that organizational or institutional support is essential for scientists to find the need for cross-disciplinary engagement and finally engage in sharing, discovering and reusing the data. One of their major concerns was the imbalance in engagement between information/IT experts and domain experts in this early stage of the EarthCube initiative: builders (information and IT experts) are more engaged than end-users or domain experts (Cutcher-Gershenfeld et al., 2016).

5 Resource Description Framework: http://www.w3.org/TR/rdf11-primer/

Roles and Opportunities for Librarians/Information Professionals

Big data curation requires concerted efforts from domain experts, information professionals, and IT experts. The typical life cycle of big data often stems from individual researchers or scientists, who are the data contributors but are not necessarily good at data management. There is an increasing demand for librarians or information professionals' role in the collaboration in the geoscientific data management. A librarian in the new data era is also an information expert or information professional. Librarians and other information professionals could play an essential role and have a vital future as a facilitator between domain experts, IT experts, and researchers. In the rapidly evolving new era of big data, information professionals must recognize the imperative of their new roles and foster new partnerships with domain experts, IT experts and researchers. They could serve as the vital intermediaries between information resources and end-users, whereas IT experts focus more on the building and maintaining of the cyberinfrastructure for the big data (Obiora Omekwu and Eteng, 2006). Information experts, to a great extent, ensure the quality and interoperability of the data and software by working in the aspects of metadata, controlled vocabularies and thesauri with input from the domain experts, geoscientists, and end-users. On the other hand, information professionals also work with IT experts to add value to the big data for long-term usage and interoperability of the linked data. Information professionals could provide guidelines or tutorials to end users to train them to enrich the descriptive information for the data record while ensuring the information is in accordance with metadata standards.

The new roles of information professionals also pose challenges during the emergence of big data. Librarians/information professionals need to acquire necessary skills, such as how semantic and linked data are used, accessed, and disseminated in real-world semantic data repositories and alliance. To play their vital roles in the big data era and add value to their work, librarians/information professionals must be prepared for the challenges of digital and IT technologies, new and novel ways of learning and research, and the demands from the end-users for data-driven research and decision making (Obiora Omekwu and Eteng, 2006). Linked data are essential for the semantic web of big data and data-driven research at present and in the future, so information professionals need to master the concept of the semantic languages and tools to query data, such as RDF and SPARQL query languages.

In the essence of collaboration and outreach, the librarian community needs to team up to outreach to the faculty and researchers in curation the big data in the academic disciplines. Librarians not only deal with books and resources in the library but expand their roles in the emerging trends and areas. The emerging "new" roles require librarians to get out from their physical setting to work with domain experts, IT experts, and researchers to add critical value to big data. This certainly requires librarians/information professionals in the new data age equipped with essential knowledge and skills for big data management and curation. The gap between information and IT experts is becoming smaller, and sometimes those two groups of experts will need to learn from each other to fully embrace the digital worlds. Librarian/information professionals also wear the hat of training end users in data input and query.

Summary

In this paper, we have provided a review of current practices in building and managing big data within the context of large data infrastructure, using geoscience cyberinfrastructure such as Interdisciplinary Earth Data Alliance (IEDA) and EarthCube as a case study, and explore the metadata/data librarian's role in big data in geoscience field. Metadata and controlled vocabulary are crucial for big data management in order to meet with challenges and opportunities for big data curation due to the growingly complex disciplinary data systems. The concerted efforts from domain experts, information professionals, IT experts are essential for the accessibility and interoperability of the big data. Furthermore, we identified the current challenges, key elements and prospects to construct a more robust and future-proof big data infrastructure for research and publication for the future, as well as the roles, requirements, and opportunities for librarians in the emerging big data era.

References

Abrams, S., Cruse, P., & Kunze, J. (2009). Preservation is not a place. *International Journal of Digital Curation 4*(1), 8–21.

Committee on Cataloging: Description and Access. (2000, June 16). Committee on Cataloging: Description & Access – Task Force on Metadata: Final Report. Retrieved from https://www.libraries.psu.edu/tas/jca/ccda/tf-meta6.html

Black, R., Katz, A., & Kretschmann, K. (2014). Earthcube: A community-driven organization for geoscience cyberinfrastructure. *Limnology and Oceanography Bulletin*, *23*(4), 80–83. doi:10.1002/lob.201423480a

Calhoun, K. (2007). Being a librarian: Metadata and metadata specialists in the twenty-first century. *Library Hi Tech*, *25*(2), 174–187. doi:10.1108/07378830710754947

Carter-Orlando, M., Ferrini, V.L., Lehnert, K., Carbotte, S.M., Richard, S.M., Morton, J.J., et al. (2017). IEDA integrated services: Improving the user experience for interdisciplinary earth science research [Abstract]. *AGU Fall Meeting*, #IN12B-03.

Cutcher-Gershenfeld, J., Baker, K., Berente, N., Carter, D., DeChurch, L., Flint, C., et al. (2016). Build it, but will they come? A geoscience cyberinfrastructure baseline analysis. *Data Science Journal, 15*(0), 8. doi:10.5334/dsj-2016-008

Davenport, T. (2014). *Big data at work: Dispelling the myths, uncovering the opportunities*. Boston, Massachusetts: Harvard Business Review Press.

DeRisi, S., Kennison, R., & Twyman, N. (2003). The what and whys of DOIs. *PLOS Biology*, *1*(2), e57. doi:10.1371/journal.pbio.0000057

Diekema, A.R., Wesolek, A., & Walters, C.D. (2014). The NSF/NIH effect: Surveying the effect of data management requirements on faculty, sponsored programs, and institutional repositories. *The Journal of Academic Librarianship*, *40*(3), 322–331. doi:10.1016/j.acalib.2014.04.010

Dutton, A., Rubin, K., McLean, N., Bowring, J., Bard, E., Edwards, R.L., et al. (2017). Data reporting standards for publication of U-series data for geochronology and timescale assessment in the earth sciences. *Quaternary Geochronology*, *39*, 142–149. doi:10.1016/j.quageo.2017.03.001

Ebert-Uphoff, I., Thompson, D.R., Demir,, I., Gel, Y. R., Hill, M.C., Karpatne, A., Guereque, M., Kumar, V., Cabral-Cano, E. and Smyth, P. (2017). A vision for the development of benchmarks to bridge geoscience and data science. *Proceedings of the Seventh International Workshop on Climate Informatics (CI 2017)*. Retrieved from https://www.engr.colostate.edu/~iebert/PAPERS/CI2017_paper_15.pdf

National Science Foundation. (2015, March 18). Today's data, tomorrow's discoveries: Increasing access to the results of research funded by the National Science Foundation. *National Science Foundation*. Retrieved from https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf

National Science Foundation. (2019). ENG data management plans. Retrieved from https://www.nsf.gov/eng/general/dmp.jsp

Gil, Y., David, C.H., Demir, I., Essawy, B.T., Fulweiler, R.W., Goodall, J.L., et al. (2016). Toward the geoscience paper of the future: Best practices for documenting and sharing research from data to software to provenance. *Earth Space Science, 3*(10), 388–415.

Guo, H. (2015). Big data for scientific research and discovery. *International Journal of Digital Earth*, *8*(1), 1–2. doi:10.1080/17538947.2015.1015942

Guo, H. (2017). Big earth data: A new frontier in Earth and information sciences. *Big Earth Data*, *1*(1-2), 4–20. doi:10.1080/20964471.2017.1403062

Haynes, D. (2018). *Metadata for Information Management and Retrieval: Understanding metadata and its use*. London, United Kingdom: Facet Publishing.

Horsburgh, J.S., Aufdenkampe, A.K., Lehnert, K.A., Mayorga, E., Hsu, L., Song, L., et al. (2014, December). Information requirements for integrating spatially discrete, feature-based earth observations [Abstract]. *AGU Fall Meeting*, *44*, IN44A–07.

Hsu, L., Martin, R.L., McElroy, B., Litwin-Miller, K., & Kim, W. (2015). Data management, sharing, and reuse in experimental geomorphology: Challenges, strategies, and scientific opportunities. *Geomorphology*, *244*, 180–189. doi:10.1016/j.geomorph.2015.03.039

Hsu, L., Lehnert, K.A., Goodwillie, A., Delano, J.W., Gill, J.B., Tivey, M.A., et al. (2015). Rescue of long-tail data from the ocean bottom to the Moon: IEDA Data Rescue Mini-Awards. *GeoResJ*, *6*, 108–114. doi:10.1016/j.grj.2015.02.012

Ji, P., Lehnert, K.A., Arko, R.A., Song, L., Hsu, L., Carter, M.R., et al. (2014). IEDA thesaurus: A controlled vocabulary for IEDA systems to advance integration. [Abstract]. *AGU Fall Meeting*, *31*, IN31D–3751.

Klump, J., Huber, R., & Diepenbroek, M. (2016). DOI for geoscience data – How early practices shape present perceptions. *Earth Science Informatics*, *9*(1), 123–136. doi:10.1007/s12145-015-0231-5

L'Abate, G., Caracciolo, C., Pesce, V., Geser, G., Protonotarios, V., & Costantini, E.A.C. (2015). Exposing vocabularies for soil as Linked Open Data. *Information Processing in Agriculture*, *2*(3), 208–216. doi:10.1016/j.inpa.2015.10.002

Lehnert, K., & Hsu, L. (2015). The new paradigm of data publication. *Elements*, *11*(5), 368–369.

Lehnert, K., Song, L., Hsu, L., Ji, P., & Carter, M. (2015, April). Interdisciplinary data resources for volcanology at IEDA (Interdisciplinary Earth Data Alliance) [Abstract]. *EGU General Assembly 2015,* id.12459.

Lehnert, K., Carbotte, S., Richard, S., Carter, M., Ferrini, V., Morton, J., et al. (2018, April). IEDA integrated services for solid earth observational data [Abstract]. *20th EGU General Assembly Conference,* p. 16860.

Lynch, C. (2008). Big data: How do your data grow? *Nature*, *455*, 28–29. doi:10.1038/455028a

Ma, X. (2017). Linked geoscience data in practice: Where W3C standards meet domain knowledge, data visualization and OGC standards. *Earth Science Informatics*, *10*(4), 429–441. doi:10.1007/s12145-017-0304-8

Miller, H.G., & Mork, P. (2013). From data to decisions: A value chain for big data. *IT Professional*, *15*(1), 57–59. doi:10.1109/MITP.2013.11

Mooney, H., & Newton, M. (2012). The anatomy of a data citation: Discovery, reuse, and credit. *Journal of Librarianship and Scholarly Communication*, *1*(1), eP1035. doi:10.7710/2162-3309.1035

Moulaison, H.L., Rathbun-Grubb, S., Abbas, J., Greenberg, J., La Barre, K., Rodríguez, E. M., et al. (2012). Emerging trends in metadata research. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–4. doi:10.1002/meet.14504901174

Nature. (2008). Community cleverness required. *Nature*, *455*(7209), 1. doi:10.1038/455001a

Obiora Omekwu, C., & Eteng, U. (2006). Roadmap to change: Emerging roles for information professionals. *Library Review*, *55*(4), 267–277. doi:10.1108/00242530610660816

Riley, J., & National Information Standards Organization (U.S.). (2017). *Understanding metadata: What is metadata, and what is it for?* Retrieved from https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf

Sweet, L.E., & Moulaison, H.L. (2013). Electronic health records data and metadata: Challenges for big data in the United States. *Big Data*, *1*(4), 245–251. doi:10.1089/big.2013.0023

Yu, L., & Liu, Y. (2015). Using linked data in a heterogeneous sensor web: Challenges, experiments and lessons learned. *International Journal of Digital Earth, 8*(1), 17–37. doi:10.1080/17538947.2013.839007

Zeng, M.L., & Qin, J. (2016). *Metadata* (2nd ed.). Chicago, Illinois: Neal-Schuman.