Review

Overview of disciplinary data sharing practices and promotion of open data in science

Jihyun Kim

Department of Library and Information Science, Ewha Womans University, Seoul, Korea

Abstract

The present study specifies the historical development of data sharing practices in three disciplines—oceanography, ecology, and genomics—along with the evolving progress of movements—e-Science, cyberinfrastructure, and open science—that expedite data sharing in more diverse disciplines. The review of these disciplinary data-sharing practices and the movements suggests opportunities and challenges that would serve as a basis for implementing data-sharing practices. The increasing need for large-scale and interdisciplinary research provides momentum for initiating data sharing. In addition, the development of data repositories and standards for metadata and data format facilitates data sharing. However, challenges need to be addressed, in regard to conflicting issues of patenting data, concerns about privacy and confidentiality, and informed consent that adequately enables data sharing. It is also necessary to consider the needs of the various stakeholders involved in data sharing to incentivize them to improve its impact.

Keywords

Data sharing; Disciplinary data sharing practices; Movements for open data; Open data; Open science

Introduction

Research data gain much attention these days due to their potential for being shared and reinterpreted from a new perspective. Data sharing is defined as "making data available to people other than those who have generated them" [1]. The modes of data sharing vary from the exchange of research data among colleagues to making data publicly accessible to others, mostly by depositing them into data repositories. The shared data are generally available for reuse, which refers to "the secondary use of data—not for its original purpose but for studying new problems" [2]. Data reuse allows researchers to reanalyze data with new perspectives and encourages greater scrutiny, which in turn leads to the advancement of science. Data sharing and

reuse also enable achieving reproducible research, allowing verification of results, and utilizing big data to solve complicated research questions [3]. Researchers expect and appreciate these benefits, which motivate their data sharing.

This study examines data sharing initiatives autonomously developed in three disciplines—oceanography, ecology, and genomics—and the movements that have promoted data sharing since 2000. By presenting a historical overview of data sharing practices and movements, the study identifies opportunities and challenges of sharing research data as well as evolving frameworks that facilitate open data in science.

Data Sharing Initiatives in Three Disciplines

There are a few disciplines that demonstrate established norms and cultures of sharing research data. The present study selected three disciplines among the fields of science in which historical reviews of data sharing practices were available. The disciplines have quite an extensive history of data sharing dating back to the 1960s and provide useful insights on best practices that are key for implementing successful policy and infrastructure.

Oceanography

There has been a long history of research in oceanography conducted internationally across disciplines. Interdisciplinary collaborations are essential in this field of study to develop profound understanding of physical, chemical, biological, and geographical phenomena in the ocean [4]. A large quantity of oceanographic data of various types have thus been gathered and accumulated by extensive research projects. In particular, the Intergovernmental Oceanographic Commission (IOC) is an institution within United Nations Educational, Scientific and Cultural Organization (UNESCO) that plays a primary role in collecting and sharing research data in this field. The IOC was established in 1960, and its mission is to promote international collaboration and coordination of research, services, and capacity-building programs to investigate nature and resources in the ocean and apply the knowledge for improving management and protection of the environment. Currently, 147 countries participate in the IOC as member states.

Since the beginning of the IOC, free and unrestricted data exchange among member states has been considered important. The emphasis on data sharing results from the fact that oceanographic data have an irreplaceable value once collected. Sea-going measurements require enormous amounts of time, effort, and resources, and thus any measurements need to be well-protected along with metadata. Furthermore, the IOC cleans and refines the measurement results into workable datasets, which member states then share for the common good [5].

Glover et al. [5] elaborated the history of data sharing in oceanography, which dates back to the 1960s when the IOC was established. Even if the early phase of computer technology and the political environment related to the Cold War hampered data sharing, the IOC organized a working group named the IOC Data and Information Exchange, which became the International Oceanographic Data Exchange (IODE), a current IOC program responsible for sharing data and information across member states. The IOC also made a general plan for the Integrated Global Ocean Station System (IGOSS), which collects and exchanges oceanographic data jointly with the World Meteorological Organization.

During the 1970s, international cooperation for collecting and exchanging oceanographic data was initiated in earnest. A multidisciplinary Marine Environmental Data and Exchange system was adopted in 1973, and it was one of the earliest metadata systems accepted in practice. The IGOSS planned in the 1960s became fully operational in 1975. The IOC also provided guidance to the General Bathymetric Chart of the Ocean, the goal of which was to facilitate scientific cooperation that supported sharing and preserving bathymetric data and associated metadata. In addition, the IOC established data centers in the US and Germany that distributed large amount of surface drifter data via the global telecommunication system of the day.

In the 1980s, the remarkable progress of computer technology enabled increased data exchange capacity. The IOC recommended a general formatting system (GF3) and made software for GF3 freely available. Data sharing was then facilitated among all institutions involved in international collaboration. The IOC also improved the functionality of the IGOSS for timely collection and exchange of standard data. Moreover, the UN Conventions of the Law of the Sea was enacted to provide a legal framework that determined international maritime communication. The law significantly affected how oceanographic data could be shared within a legal boundary.

In the 1990s, the IOC embraced global programs such as the World Meteorological Organization and UN Environment Programme to ensure cooperation for organizing the Global Ocean Observing System, which later evolved into the Global Climate Observing System. In addition, the IODE Global Oceanographic Data Archaeology and Rescue project was launched in the late 1990s. All the data from the project were disseminated on DVD and uploaded to the World Ocean Database online. Improved computer technology and widespread use of the Internet in the late 1990s also allowed for developing countries to become involved in the Ocean Data and Information Network, through which the IODE expanded services for data exchange.

From 2000 to 2010, the IOC supported greater international collaboration and open data. In 2003, the IOC Oceanographic Data Exchange Policy was announced and specified timely, free, and unrestricted sharing of data collected under IOC sponsorship as well as associated metadata and all derivative products. The IOC has also sponsored projects regarding ocean carbon science and observations as well as a related project, called the Surface Ocean CO_2 Atlas, which attempted to establish a standard data format that facilitated making data publicly available. Furthermore, the Open Data Portal was developed in 2007 for seamless access to all oceanographic data on the IODE network.

Ecology

Prior to 1950, most research projects in ecology were carried out by a small number of scientists with limited funding. This small-scale research tradition was transformed by the emergence of big ecology, which represented interdisciplinary research performed with an international scope [6]. The large-scale projects that led to big ecology established policies and guidelines for data sharing and management, and this effort helped data sharing to become an accepted norm in ecology.

Michener [7] described characteristics of large ecological projects since the 1960s based on the historical analysis of ecological collaborations conducted by Coleman [6]. The International Biological Program (IBP) was an early collaborative project that examined a broad range of biomes in a multidisciplinary scope. It was successful in ecological data synthesis and the adoption of holistic approaches to ecosystems. However, data policies and protocols were not uniformly identified, and therefore IBP data were not systematically managed. As a result, it was almost impossible to discover and acquire IBP data at this time.

The US Long-Term Ecological Research (LTER) began in 1980 and has been implemented until the present with more than 24 sites in the US territories and Antarctica. In the 1980s, data managers were hired for several sites, although data were used only by data collectors and their collaborators. Making data more widely available was realized in the 1990s due to two changes in data management practices. First, a data catalog describing core data sets in all LTER sites was published. It helped identify what data were available and where they were located. Second, the first formal guideline that required each site to establish a data management policy was provided. The guideline specified roles and responsibilities of data contributors and users and recommended providing metadata, preserving data for the long-term, and making them available in a timely manner.

In 1993, the International LTER was organized and has been expanded to include 40-member networks. The LTER networks adopted a network-wide policy in 1997, and a new data sharing policy was enacted in 2005, which strengthened the 1997 policy by defining responsibilities of data collectors and determining data embargo periods to be no more than two years after the data were collected. In addition to the adoption of the policy, the LTER established the Ecological Metadata Language, a metadata content standard that improved access to data and metadata on the LTER networks.

The National Center for Ecological Analysis and Synthesis (NCEAS) was an innovative project established in 1995; a working group consisting of 8 to 15 scientists brought existing data and collaborated on synthesizing those data and information. The NCEAS developed an informatics staff that helped the working group to manipulate and analyze data brought to the center. The NCEAS also played an important role in developing a metadata management software called Morph and establishing a data repository called the Knowledge Network for Biocomplexity, where working group members and others deposited ecological and related data.

Since 2000, several notable developments have been made for sharing biodiversity and ecological data. The Global Biodiversity Information Facility was launched in 2001, and it developed the global data portal in 2007 to promote public access to biodiversity data. Currently, the National Ecological Observatory Network and the Ocean Observatories Initiative funded by the US National Science Foundation are two major environmental observatories that provide open access to ecological data collected from territorial, freshwater, ocean, and coastal sites.

Genomics

Genomics is a field of study that examines the "whole genomes of organisms" and "uses a combination of recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the structure and function of genomes" [8]. Researchers in genomics intensively utilize instruments for mapping and sequencing nucleic acid and generate large data sets including DNA sequences, genomic locations, and functional analyses of genes and proteins. As research projects in this field become larger in scale, the need to share data is more pressing than ever before since no single laboratory can fully investigate the overwhelmingly large amounts of data.

Cook-Deegan et al. [9] discussed the historical aspects of research projects that contributed to facilitating data sharing in genomics. The Human Genome Project (HGP) was the first and foremost important project due to not only the achievements of generating a human genome reference sequence and developing new technologies and instruments, but also the establishment of principles and policies for sharing DNA se-

quences created by the project. The idea of the HGP was originally brought up in 1985, with a focus on creating a reference sequence of the human genome as a tool for research and application. By early 1996, the HGP decided to accelerate large-scale human genome sequencing with five national partners and had a meeting in Bermuda.

In this meeting, participants announced the Bermuda Principles, which replaced previous guidelines for data sharing that applied only to the National Institute of Health and the US Department of Energy. The former guidelines also required sharing data from DNA mapping and sequencing within six months of generation. The Bermuda Principles, however, strongly recommended daily releases of all HGP-generated DNA sequences. The principles provided a project-wide policy for the first time, which helped develop the Human Sequence and Mapping Index, a website that identified laboratories for DNA sequencing distributed around the world and avoided duplication of research.

The Bermuda meeting in 1996 also raised the issue of patenting DNA molecules and methods, which became common at that time. The majority of Bermuda attendees strongly disagreed with the patents and attempted to develop a data sharing policy that clearly identified the purpose: "All human genomic sequence information…should be freely available and in the public domain in order to encourage further research and development, and to maximise its benefit to society" [10]. In 1998, the Bermuda Principles became an official data sharing policy of the HGP that applied to any genome sequence data publicly funded by participating countries—the US, the UK, France, Germany, Japan, and China.

The HGP generated a human genome reference sequence between 2000 and 2003. Publicly-funded HGP sequences from laboratories in the six countries entered the public domain in 2001, and access to the data became open and free via Genbank. However, genome sequences generated from a private company named Celera Genomics, a participant in the HGP, were not completely open to the public. The data were available for free for non-commercial use in parcels of 1 kb or otherwise based on subscription fees or an access agreement with the company. The two modes of data sharing raised intense controversy over access to sequence data. A policy was later developed that recommended sharing data accompanying publications rather than the original Bermuda Principles recommending daily releases of DNA sequences when they were generated.

In 2003, when the HGP was nearly ended, the Wellcome Trust organized a meeting in Fort Lauderdale, Florida to define several important statements regarding data sharing in genomics. The meeting focused on the value of daily releases of DNA sequences, which followed the ethos of the Bermuda

Principles. At the same time, it mandated that data generators must be credited when data were reused. In addition, another genome sequence repository named dbGAP was developed in 2006. Unlike Genbank, dbGAP has two tiers, one that provides publicly available data and the other that collects private data with identifiable information. The latter requires data access committees to ensure that users have appropriate reasons to access the data. In 2007, dbGAP was designated as a primary repository for depositing data from the National Institute of Health's Genome-Wide Association Studies. The Genome-Wide Association Studies data sharing policy allows for a six-month embargo of data for validation, and dbGAP keeps the data private during the embargo period. After six months, the data are freely available to the public. The adaptations of data sharing policies and data repositories indicated that the rights and interests of both researchers and human subjects should be considered.

A recent development in the Global Alliance of Genomics was the creation of a global commons of genomics that started with 50 individuals from eight countries in 2013. It now contains around 500 institutional members from 71 countries. One of the outstanding projects that the alliance made was the Breast Cancer Gene (BRCA) Challenge, a database that curated variants of the two most studied and clinically important genes, BRCA1 and BRCA2. These are tumor suppressor genes, and some mutations of the genes are known to cause breast, ovarian, and other cancers.

The BRCA Challenge consisted of three tiers. The first tier was entirely public and provided variants with interpretations made by experts in the field. The second tier provided evidence-based research data, including conflicting interpretations of variants and reports. The third tier included case-level data linked to identifiable individuals and thus required the highest level of security. The BRCA Challenge intended to open BRCA data against patented data related to BRCA studies, which helped scientists to catch up on genomic research.

The historical development of data sharing practices in the three disciplines illustrates opportunities and challenges for data sharing. First, rapidly developed computing power and encouragement of international and cross-disciplinary collaboration were the primary impetus for active data sharing in the disciplines. Second, the development of metadata, formatting standards, and data repositories to archive and access data further facilitated data sharing. Third, a considerable conflict of interest was identified specifically in genomics between researchers who supported open data and those who agreed with patenting DNA sequences. This implies disagreements that researchers might have regarding incentives and disincentives of data sharing. Different layers in genome databases with distinct levels of access to data also indicate the

need for privacy considerations, confidentiality of human subjects, and appropriate informed consent procedures.

Movements that Support Open Data in Science

The disciplines that already have a long history of data sharing practices indicate that norms for data sharing have been established. Yet, the culture of sharing data is prevalent only in certain disciplines. Data sharing has been facilitated in a wider range of disciplines by movements that promote the emergence of data-centric science and the importance of data sharing. The following sections describe influential movements since 2000, including e-Science, cyberinfrastructure, and open science.

e-Science and cyberinfrastructure

The term e-Science was originally proposed in 1999 by John Taylor, Director General of Research Councils in the UK Office of Science and Technology. Predicated on his experience as the head of Hewlett-Packard Lab, he recognized the transformation of scientific research as a new way of collaborative, interdisciplinary, and data-intensive work. In this sense, he defined e-Science as "global collaboration in key areas of science and the next generation of infrastructure that will enable it" [11]. Hey and Trefethen [12] further elaborated that e-Science represents large-scale and highly complex scientific problems for which efforts by distributed, collaborative, and multi-disciplinary teams are needed as well as the collaborative tools and technologies required to solve these problems. In this e-Science environment, it is evident that enormous amounts of research data and metadata were being generated and accumulated, and thus how to manage the data deluge became an important issue. Data repositories thus played a significant role in e-Science infrastructure by not only preserving the wealth of scientific data but also providing programs to manipulate and visualize them [13].

Sharing data generated from government-funded research was also necessary in order to fulfill the premise of e-Science. The roles and responsibilities of governments were emphasized to address data sharing issues beyond national boundaries and to build international cooperation for issues of global significance. As a result of these discussions, 13 Organization for Economic Cooperation and Development (OECD) principles and guidelines for access to data from public funding were proposed in 2007: openness, flexibility, transparency, legal conformity, protection of intellectual property, formal responsibility, professionalism, interoperability, quality, security, efficiency, accountability, and sustainability. The principles implied that data from public research should be available as widely as possible, with consideration for legal and ethical

conditions and based on efficient and accountable data management.

Cyberinfrastructure was a term first introduced in 2003 by the US National Science Foundation's Blue Ribbon Advisory Panel on Cyberinfrastructure, which defined it as "infrastructure based upon distributed computer, information and communication technology" [14]. The underlying layer of cyberinfrastructure consisted of technological components related to computation, storage, and communication. The upper layer was composed of software programs, services, instruments, data, information, knowledge, social practices, and communities of practices. Cyberinfrastructure between the two layers "should provide an effective and efficient platform for the empowerment of specific communities of researchers to innovate and eventually revolutionize what they do, how they do it and who participates" [14].

Both e-Science and cyberinfrastructure signified computational infrastructure that enabled researchers to assemble heterogeneous data and information sources as well as to make scientific analyses and visualizations based on a substantial quantity of data. Large research grants were offered for research and technical development of cyberinfrastructure, which led to a new form of investigation and cross-disciplinary collaboration [15].

The scope of openness in e-Science was also discussed since the commitment to disclosing research outputs varies by disciplines, although it was framed in science as a whole. It was suggested that the degree of openness towards research materials perceived by scientists ranged from private management through peer exchange to public sharing. The various perceptions of openness imply that it was necessary to understand the benefits that researchers expected or actually gained and how it affected their patterns of providing open access to research materials, methods, tools, and resources [16].

Open science

Open science is a movement that evolved through previous efforts to provide open access to research results, increasing motivations to share resources between disciplines, and the need for greater efficiency, accountability, and reproducibility of research [17]. The 2015 OECD report suggests that open science indicated efforts by various stakeholders in scientific communities, including researchers, governments, and funding agencies, to make publications and research data publicly accessible in digital format with no or minimal restrictions [18]. This notion of open science, however, is regarded as narrow, and a recent definition of open science encompasses a wide range of activities in science: "Open science is the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research

processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods" [19]. The European Commission has an even broader definition of open science, which it describes as "a new approach to the scientific process based on cooperative work and new ways of diffusing knowledge by using digital technologies and new collaborative tools" [20]. This definition represents a shift from the traditional way of creating publications to sharing and using all available knowledge as early as possible during the research process.

Open science is an overarching term that concerns various movements for sharing publications, data, methods, resources, and tools at any stage of the research process. Fecher and Friesike [21] determined the movements involved in open science as five schools of thoughts: democratic, infrastructure, measurement, pragmatic, and public. The democratic school focuses on making research products available and has two main streams, open access to publications and open data. The infrastructure school involves building technological infrastructure, mostly software applications that enable research via the internet. The measurement school concerns developing alternative standards and measures to assure scientific impact, known as altmetrics, which deal with other forms of publication and social media as a source of scientific contribution. The pragmatic school considers open science as a method for making research and knowledge dissemination more efficient and believes that science can be advanced by opening and reinventing knowledge production processes. Lastly, the public school involves making science accessible to the public and suggests that scientists increase accessibility of research processes and research results for citizens.

In particular, open data is one of the core topics in open science and evolved from the OECD principles and guidelines of 2007. An increasing number of government funding agencies are adopting these principles and developing policies that strongly recommend data sharing. For instance, the National Science Foundation has specified the requirement of a data management plan for all grant proposals since 2011. A data management plan must include types of data and other materials from a proposed research, standards for data and metadata, policies for data sharing and reuse, and plans for archiving data [22]. In addition, scholarly journals have gradually enacted data sharing policies. Kim and Stanton [23] found that journals' enforcement of their data policies positively affected researchers' data sharing.

Conclusion

The present study describes the historical development of data sharing initiatives in oceanography, ecology, and genom-

ics as well as movements facilitating data sharing, which have evolved since 2000 from e-Science and cyberinfrastructure to open science. The review shows various social and technological activities that have promoted data sharing as well as conflicting issues that indicate restrictions for open data. Specifically, open science encourages researchers to disclose not only publications and data, but also methods, tools, and resources for the advancement of science. It is important to consider the different needs of stakeholders—researchers, study participants, governments, funding agencies, journals, and publishers—when developing policies and procedures, which will improve the impact of data sharing.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

References

1. US Department of Energy. EERE digital data management glossary [Internet]. Washington, DC; Energy Efficiency & Renewable Energy [cited 2019 Feb 11]. Available from: https://www.energy.gov/eere/funding/eere-digital-data-management-glossary
2. Yoon A. Data reusers' trust development. J Assoc Inf Sci Technol 2017;68:946-56. https://doi.org/10.1002/asi.23730
3. National Academy of Science, Engineering and Medicine. Benefits of data sharing. In: National Academy of Science, Engineering and Medicine. Principles and obstacles for sharing data from environmental health research: workshop summary. Washington, DC: National Academy Press; 2016. p. 29-38.
4. Powell TM. The rise of interdisciplinary oceanography. Oceanography 2008;21:54-7. https://doi.org/10.5670/oceanog.2008.35
5. Glover DM, Wiebe PH, Chandler CL, Levitus S. IOC contributions to international, interdisciplinary open data sharing. Oceanography 2010;23:140-51. https://doi.org/10.5670/oceanog.2010.29
6. Coleman DC. Big ecology: the emergence of ecosystem science. Berkeley: University of California Press; 2010.
7. Michener WK. Ecological data sharing. Ecol Inf 2015;29:33-44. https://doi.org/10.1016/j.ecoinf.2015.06.010
8. EMBL-EBI. What is genomics? [Internet]. Cambridgeshire: EMBL-EBI [cited 2019 Feb 11]. Available from: https://www.ebi.ac.uk/training/online/course/genomics-introduction-ebi-resources/what-genomics
9. Cook-Deegan R, Ankeny RA, Maxson Jones K. Sharing data to build a medical information commons: from Ber-

muda to the Global Alliance. Annu Rev Genomics Hum Genet 2017;18:389-415. https://doi.org/10.1146/annurev-genom-083115-022515

10. Wellcome Trust. Funding guidance: statement on genome data release [Internet]. London: Wellcome Trust [cited 2019 Feb 11]. Available from: https://wellcome.ac.uk/funding/guidance/statement-genome-data-release

11. Hey T, Trefethen A. e-Science and its implications. Phil Trans R Soc London A Math Phys Eng Sci 2003;361:1809-25. https://doi.org/10.1098/rsta.2003.1224

12. Hey T, Trefethen A. E-science, cyber-infrastructure, and scholarly communication. In: Olson GM, Zimmerman A, Bos N. Scientific collaboration on the Internet. Cambridge, MA: MIT Press; 2008. p. 15-33.

13. Hey T, Trefethen A. The data deluge: an e-science perspective. In: Berman F, Fox G, Hey T. Grid computing: making the global infrastructure a reality. New York, NY: John Wiley & Sons; 2003. p. 809-24.

14. Atkins D, Droegemeier KK, Feldman SI, et al. Revolutionizing science and engineering through cyberinfrastructure: report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure [Internet]. Alexandria, VA: National Science Foundation; 2003 [cited 2019 Feb 11]. Available from: https://www.nsf.gov/cise/sci/reports/atkins.pdf

15. Pacheco RC, Nascimento ER, Weber RO. Digital science: cyberinfrastructure, e-Science and citizen science. In: North K, Maier R, Haas O, editors. Knowledge management in digital change. [place unknown]: Springer; 2018. p. 377-88.

16. Whyte A, Pryor G. Open science in practice: researcher perspectives and participation. Int J Digit Curation 2011;6:199-213.

17. Bezjak S, Clyburne-Sherin A, Conzett P, et al. Open science training handbook [Internet]. Geneva: Zenodo; 2018 [cited 2019 Feb 11]. Available from: https://doi.org/10.5281/zenodo.1212496

18. Organization for Economic Cooperation and Development. Making open science a reality [Internet]. Paris: Organization for Economic Cooperation and Development; 2015 [cited 2019 Feb 11]. Available from: http://dx.doi.org/10.1787/5jrs2f963zs1-en

19. FOSTER. Open science definition [Internet]. [place unknow]: FOSTER [cited 2019 Feb 11]. Available from: https://www.fosteropenscience.eu/foster-taxonomy/open-science-definition

20. European Commission. Open innovation, open science and open to the world. Luxembourg: Publications Office of the European Union; 2016.

21. Fecher B, Friesike S. Open science: one term, five schools of thought. In: Fecher B, Friesike S. Opening science. [place unknown]: Springer; 2014. p. 17-47.

22. Pasek JE. Historical development and key issues of data management plan requirements for National Science Foundation grants: a review. Issues Sci Technol Librariansh 2017 Summer [Epub]. https://doi.org/10.5062/F4QC01RP

23. Kim Y, Stanton JM. Institutional and individual factors affecting scientists' data-sharing behaviors: a multilevel analysis. J Assoc Inf Sci Technol 2016;67:776-99. https://doi.org/10.1002/asi.23424