Chapter 74
Big Data Sharing Among Academics

Jeonghyun Kim
University of North Texas, USA

ABSTRACT

The goal of this chapter is to explore the practice of big data sharing among academics and issues related to this sharing. The first part of the chapter reviews literature on big data sharing practices using current technology. The second part presents case studies on disciplinary data repositories in terms of their requirements and policies. It describes and compares such requirements and policies at disciplinary repositories in three areas: Dryad for life science, Interuniversity Consortium for Political and Social Research (ICPSR) for social science, and the National Oceanographic Data Center (NODC) for physical science.

INTRODUCTION

The September 2009 issue of *Nature* included an interesting special section on data sharing. An opinion article in the section discussed the Toronto International Data Release Workshop, where attendees "[recommended] extending the practice to other biological data sets" (Birney et al., 2009, p. 168) and developing a set of suggested best practices for funding agencies, scientists, and journal editors. The February 2011 issue of *Science* compiled several interesting articles to provide a broad look at the challenges and opportunities posed by the data deluge in various areas of research, including neuroscience, ecology, health, and social science, where there is a demand for the acquisition, integration, and exchange of vast amounts of research data.

The term *big data* is a current buzzword. It is a loosely defined term to describe massive and complex data sets largely generated from recent and unprecedented advancements in data recording and storage technology (Diebold, 2003). Explosive growth means that revolutionary measures are needed for data management, analysis, and accessibility. Along with this growth, the emergence of a new "fourth paradigm" (Gray, 2009) for scientific research, where "all of the science literature is online, all of the science data is online, and they interoperate with each other" (Howe et al., 2008, p.

DOI: 10.4018/978-1-4666-9840-6.ch074

47), has created many opportunities. Therefore, the activity of organizing, representing, and making data accessible to both humans and computers has become an essential part of research and discovery.

Given the significance of this context, data sharing has become a hot topic in the scientific community. Data is a classic example of a public good in that shared data do not diminish in value. In particular, scientific data have long underpinned the cycle of discovery and are the dominant vehicles by which scientists earn credit for their work. So shared data have served as a benchmark that allows others to study and refine methods of analysis, and once collected, they can be creatively repurposed indefinitely by many hands and in many ways (Vision, 2010). Sharing data not only reinforces open scientific inquiry but also promotes new research and expedites further discovery (Fienberg, 1994). As science has become more data intensive and collaborative, data sharing has become more important.

Promoting the effective sharing of data is an increasing part of national and international scientific discourse and essential to the future of science (National Science and Technology Council, 2009). Today, many U.S. government agencies recognize that scientific, biomedical, and engineering research communities are undergoing a profound transformation in regard to access to and reuse of large-scale and diverse data sets; as such, these agencies have developed policies that mandate and/or encourage data sharing. For instance, the National Science Foundation (NSF) expects grantees to share their primary data, samples, physical collections, and other supporting materials created or gathered in the course of work under the grant.[1] The National Institutes of Health (NIH) has had a data-sharing policy since 2003; the policy states that any investigator submitting a grant application seeking direct costs of $500,000 or more in any single year is expected to include a plan to address data sharing in the application or state why data sharing is not possible.[2]

To support these needs, infrastructure is being built to store and share data for researchers as well as educators and the general public. In 2008, the NSF awarded nearly $100 million over 5 years to data preservation and infrastructure development projects under the DataNet initiative.[3] DataONE[4] is one of the awards, which is dedicated to large-scale preservation and access to multiscale, multidiscipline, and multinational data in biology, ecology, and the environmental sciences. Recently, the White House announced a $200 million initiative to create tools to improve scientific research by making sense of the huge amount of data now available. Programs like these are needed to improve the technology required to work with large and complex sets of digital data.[5]

Researchers and scientists in academia, industry, and government may choose to store and share their data in a number of ways. Among the various means, data repositories often appear to offer the best method of ensuring that data are preserved and presented in a high-quality manner and made available to the largest number of people. Data repositories are constructed with the chief goal of storage and preservation and emphasize use/reuse. In other words, the implementation of data repositories is constrained by not only the needs of data sharing but also concurrent data access. They have data as its primary focus and are often shared by a scientific community.

The goal of this chapter is to explore the practice of big data sharing among academics and issues related to this sharing. The background section of this chapter reviews literature on researchers' practices and trends with regard to data sharing and access. The main section reviews disciplinary data repositories in the areas of social science, life science, and physical science, and describes and compares the requirements and policies at disciplinary repositories. It also examines recommended and accepted file formats and data structure repositories, metadata, and specifications and guidelines on data access and sharing.