Theoretical Study of Data Diversity

Authors: Lu Cainü, Gu Liping, Nie Hua, Gu Liping

Date: 2021-11-24T00:00:00+00:00

Abstract

Data diversity constitutes an essential attribute of data. In the context of rapid advances in information technology and the open scientific data movement, the characteristics of data diversity have become increasingly pronounced. This paper first elaborates in detail on the internal and external manifestations of data diversity. Internal manifestations encompass: different entities involved in the scientific data production process, the tripartite nature of data publishing, and varying data formats employed by different disciplines during data collection and temporary storage. External manifestations include: the data lifecycle accelerating data diversity, the research lifecycle augmenting data diversity, and diversity emerging from data being shaped during specific applications. Subsequently, the article briefly introduces the common characteristics and influencing factors of data diversity, and delineates its application representations from three perspectives. For libraries and librarians, comprehending data diversity can assist researchers in addressing data submission tasks and data disclosure pressures, thereby facilitating data reuse and conforming to an ideal data ecosystem. Consequently, as a data librarian, one must possess data management competencies and understand relevant laws, regulations, policies, and agreements pertaining to data ethics, endeavoring to provide data value-added services for researchers.

Full Text

Preamble

Theoretical Research on Data Diversity

Lu Cainü[1], Gu Liping[2],[3], Nie Hua[4]
([1] Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203;
[2] National Science Library, Chinese Academy of Sciences, Beijing, 100190;
[3] Department of Library, Information and Archives Management, University of Chinese Academy of Sciences, Beijing, 101408;
[4] Peking University Library, Beijing, 100190)

Abstract

Diversity is the essential attribute of data, especially scientific data. In the context of rapid development of information technologies (ITs) and the era of open research data, the characteristics of data diversity have become increasingly evident. This paper first elaborates on the internal and external manifestations of data diversity. The internal manifestations include: different objects in the scientific data production process, the trinity of data publishing, and different data formats when collecting and depositing data across various disciplines. The external manifestations include the data curation lifecycle accelerating data diversity, the research lifecycle increasing data diversity, and diversity emerging from data being shaped during practical application. Subsequently, the paper describes the common features and impact factors of data diversity, and introduces the application representation of data diversity from three aspects. For libraries and data librarians, recognizing data diversity can help researchers address data deposit requirements and data disclosure pressures, making data reuse simpler and aligning with an ideal data ecosystem. Therefore, as a data librarian, one needs data management capabilities and knowledge of relevant laws, regulations, policies, and agreements concerning data ethics, striving to provide data value-added services to researchers.

Keywords: Diversity of data; Scientific data; Research data; Data services

Classification Number: G350

1 Introduction

Data diversity possesses three internal and three external characteristics. The internal characteristics are static, inherent properties of data as an object or entity. The external characteristics are dynamic, emerging from interactions between data and its environment and users. Data diversity has always existed but remained undiscovered and underdeveloped primarily because disciplines were relatively independent with limited cross-disciplinary interaction, and data mainly served as an adjunct to research work. However, in the era of data-driven research, the value and status of data have been continuously excavated and emphasized, making the issue of data diversity increasingly prominent. Ignoring data diversity will hinder the production and management of new data. As we transition from data exchange to data openness, top-down policy guidance requires unified management, deposit, and even open sharing of scientific data, giving rise to a series of contradictions and problems that further highlight the characteristics of data diversity. The formulation and evolution of social systems and policies related to data security and data trading, particularly in China, have also promoted the development of data diversity. This final reason is especially important because while the United States and Europe have not yet clearly recognized data diversity, China' s library community has long been aware of it, though it has failed to form a clear concept and definition until one of the authors participated in research on bibliographic diversity and gained

sudden insight.

In April 2020, the Central Committee of the Communist Party of China and the State Council issued the "Opinions on Building a More Complete System and Mechanism for Market-oriented Allocation of Factors," officially recognizing "data" as a new type of production factor and explicitly proposing strategies to accelerate the cultivation of data factor markets. The market-oriented allocation of data factors has risen to a national strategy, further highlighting its importance. In October 2021, Xi Jinping emphasized during the 34th collective study session of the Political Bureau of the CPC Central Committee the need to grasp the development trends and patterns of the digital economy and promote its healthy development in China. Data assetization, data products, and data services will be important forces driving the future development of data factor markets.

2 The Birth of Open Research Data Sharing and Data Diversity Theory

The concept of open access to research data can be traced back to the 1950s, but it has only attracted widespread attention in the last decade. The 2003 Berlin Declaration recognized research data as part of academic knowledge and required open access. Since then, stakeholders including national government agencies, research institutions, funding bodies, and academic publishers worldwide have successively formulated policies for open sharing of research data. As an increasing number of academic journals require open sharing of underlying data and the number of data journals or hybrid journals publishing data papers continues to grow, research data is no longer merely a byproduct or adjunct of research activities but has gradually become one of the main products of scientific research. Meanwhile, information technology provides technical support for research data sharing in terms of storage, transmission, and processing, and continuous technological development has accelerated the pace of research data sharing. In its 2019 report "The Future of Research: Drivers and Scenarios for the Next Decade," Elsevier pointed out that open sharing of research data based on information technology development will become the most significant feature of research activities in the next decade, potentially triggering major transformations in research organization and innovation models.

One inevitable result of open research data sharing has been the emergence of the concept and theory of data diversity. At the national level, data resources have become or are becoming a new type of production factor, with major countries worldwide competing to seize the high ground of research data open sharing to 争夺 global research data resources or protect their own from being harvested. From the perspective of research institutions, funding agencies, academic publishers, and other stakeholders, active participation in research data sharing helps enhance their influence and voice. For the research and academic ecosystem, open sharing of research data enables verification of research results, enriches the subject matter of academic publishing processes, and promotes a

virtuous cycle in the academic publishing ecosystem.

Under the trend of research data sharing, the diversity of subjects and growth in numbers among data producers, managers, and users have become increasingly apparent. Data diversity was originally only a problem to be faced in specific research work or data processing tasks. However, in the open science and open data environment, when we need to formulate corresponding sharing norms and principles or various laws, regulations, and policies, or provide information services, data services, and knowledge services based on scientific data, we need a framework agreed upon (and recognized) by all parties to consider and view data. At this point, the issue of data diversity inevitably emerges and becomes prominent.

3.1.1 Internal Manifestations of Data Diversity

The internal manifestations of data diversity include several aspects. First, data inherently possesses diverse characteristics. Across different disciplines, datasets take various forms: social sciences frequently use spreadsheet data with variables and values; life sciences often employ coded data describing organizational structures; physical sciences utilize modeling data for computer simulations; and scientific disciplines focused on observational records primarily use digital images and voice recordings. For example, as shown in [Figure 1: see original paper], environmental field measurement data, microbial sequencing data, protein sequence data, and high dynamic range image data are all distinct from one another when compared.

Second, the trinity of dataset, data description, and metadata in research data publishing represents another internal manifestation of data diversity, emphasizing the growth and development of data itself. Here, (1) the dataset, also called data entity, serves as evidence for reproducing research results; (2) the data description is a document explaining the data collection instruments, methods, generation process, funders, etc.; and (3) the metadata is information describing the dataset' s contributors (or producers), affiliated institutions, disciplines, dates, versions, and other attributes. The field of metadata cataloging also reflects data diversity when using the new RDA to describe books or other entities —the same RDA offers different options during cataloging, such as using local controlled vocabularies instead of those provided in the toolkit, thereby making cataloging data extensible.

Finally, from the perspective of computer data processing, data itself transforms into multiple formats due to various practical needs. For instance, the same set of proteomics data can be represented and stored using sequences, dictionaries, tuples, lists, etc.; the same substance can also be represented and stored using two-dimensional or three-dimensional structural diagrams, molecular formulas, simplified structural formulas, electron configurations, etc. This represents another internal manifestation of data diversity.

3.1.2 External Manifestations of Data Diversity

The external manifestations of data diversity primarily emphasize the extensibility, expandability, and applicability of data diversity. The external manifestations also have three aspects. First, during its lifecycle or when being produced or reused, data itself generates multiple different versions, each of which more or less incorporates new datasets or extracts subsets from them—this is also a form of diversity. Different datasets and different data versions constitute the first external characteristic of data diversity. In essence, the data lifecycle accelerates data diversity, or rather, the data lifecycle is a process of continuously accelerating data diversity.

Second, the research lifecycle increases data diversity. In the continuously cyclical and ascending research lifecycle—understanding trends, generating ideas, designing and organizing projects, designing experiments, conducting experiments, collecting and organizing data, analyzing data, communicating and publishing, and preserving results—different people and different producers generate different data, and experiments conducted at different times also produce different data. In other words, the internal and external environments of data produce diverse data, with each cycle generating different data. For example, different data produced at various processing stages include: raw data collected by instruments, derived data extracted or merged, and research data selected for validation results. Additionally, if we consider all academic records involved in the broad research lifecycle as a type of data and manage and preserve them, the data itself is also different. As information and data have become ubiquitous forms embedded in social production, life, and consumption processes, the connotation and boundaries of information resources have expanded from documentary information to data, literature, entities, and all other forms of digital existence. At this point, data has become a broad and generalized concept, and diversity has consequently become prominent.

Finally, in specific application fields such as data science, big data, artificial intelligence, data modeling, and intelligent data, data inevitably changes to adapt to different machines and application/software requirements—it must be shaped and reshaped. This process causes another change in data diversity, including changes in storage formats to accommodate different machines or software.

The internal and external manifestations of data diversity are not isolated aspects but two sides of the same coin, complementing each other. First, the attribute values in metadata represent another form of external manifestation. The clearer and more standardized the internal manifestations of data diversity, the greater its potential for application, and the more evident its external manifestations become. Second, from the perspective of all research data as a whole, the richer the external manifestations of data diversity, the more fertile and healthy the soil and environment for data production become, which to some extent drives the generation of more datasets and their descriptions and metadata, making the internal manifestations of data diversity more complete,

unified, and diversified.

Data diversity constitutes the various internal and external manifestations formed during the research process and in the open science data environment. It is one of the essential attributes of data, with the purpose of continuously self-growing and self-developing to achieve efficient governance and application of data. It can be said that data diversity is both a means and an end.

3.2.1 Subject Diversification

Subject diversification is one of the main characteristics of data diversity. From the perspective of the research data lifecycle, subject diversification is reflected in almost all processes of the data lifecycle. For example, in the data generation stage, data producers are diversified—not only in terms of large numbers and wide distribution but also in diverse types, which may be observation machines, computers, researchers, or laboratory technicians. In the data storage stage, data may be stored in diverse systems, including institutional repositories, public data repositories, or personal computers. From the internal manifestation perspective, the maintenance subjects of metadata, data descriptions, and datasets are also diversified. The metadata subject may be libraries and librarians or storage personnel, while the subjects of data descriptions and datasets may be researchers or observation machines.

3.2.2 Collaborative Development

Another characteristic of data diversity is collaborative development among different subjects. After researchers produce datasets, data management (or governance) personnel need to process and handle data descriptions and metadata. Subsequently, analysts may only select partial data or subsets for analysis and visualization, and users may also generate new data or data descriptions based on this research data, as well as traders or trading platforms responsible for data commodity transactions that are forming or may emerge in the future to sell data products. It is evident that in the data lifecycle or research data ecosystem, subjects playing or exerting different roles collaborate and co-govern in a diversified manner, jointly promoting research data diversity.

3.2.3 Formulation of Common Rules

To achieve data diversity, especially to promote the unification and completeness of its internal manifestations, different stakeholders in the scientific, publishing, and library communities should formulate common rules based on mutual collaboration to promote long-term governance and efficient utilization of data. Currently, a series of rules and standards related to research data have been introduced internationally. For example, in terms of metadata standards, there are approximately 65 research data metadata standards worldwide, with common ones including Dublin Core, Data Documentation Initiative (DDI), Ecological Metadata Language (EML), ISO 19115, and FGDC-CSDGM for the geospatial

domain. For data management and sharing, there are the FAIR data sharing principles recognized by numerous global organizations and institutions. For data citation, data publishing and storage systems strive to provide persistent identifiers (PID) or Digital Object Identifiers (DOI) for each piece of data, along with data citation principles and standards. Currently, the Research Data Alliance (RDA) and the World Data System (WDS) have jointly established the Scholarly Link Exchange Working Group to formulate linking rules between papers and research data and provide services. The National Information Standards Organization (NISO) has also announced the launch of a new project to link workflows between publishers and repositories, achieving mutual linking between research data and papers, and forming a series of standards or best practices for metadata, terminology, and citation/linking types for data-paper relationships.

4 Influencing Factors of Data Diversity

As previously stated, data diversity is both a means and an end. So how can we achieve this goal of data diversity? Or how do we drive and maintain data diversity? This paper briefly elaborates from the following three aspects.

4.1 Environmental Factors

A series of environments led by the natural environment, including political, economic, legal, and technological environments, form the foundation of human survival and life. To a certain extent, these environmental factors are also the main influencing factors for data diversity. In particular, open sharing environments and various well-developed legal, technological, and economic environments are all fundamental factors for diversified data production and use. Additionally, competitive environments, including competition at the individual, institutional, and national levels, are also important factors for data diversity. Without competition, monopolies may emerge, which would certainly not be conducive to data diversity.

4.2 Technical Means

In the information and digital age, data is typically stored and displayed in digital form, which cannot be separated from databases, networks, and information communication technologies (collectively referred to as information technology). In the artificial intelligence era, AI technology can influence data diversity subset recommendations, big data processing and analysis, high-speed data storage and transmission, thereby affecting data diversity. Furthermore, blockchain, cloud computing, and other technologies also affect data storage and transmission, while data analysis and visualization technologies affect data application and display. All these technologies influence data diversity from various levels and perspectives, especially its external manifestations.

4.3 Standard Compliance

Without compliance with standards and norms, data would not only be diverse but would become chaotic and disorderly, even to the point of being undiscoverable, inaccessible, unobtainable, and unusable. A series of standards related to research data, including data publishing standards, data citation standards, metadata standards, data description standards, data use standards, and possible future research data-paper linking standards, are all safeguards for orderly diversification under data rules. If all stakeholders in the data ecosystem (including producers, publishers, managers, users, funders, etc.) do not comply with various standards, the rights and interests of data producers cannot be guaranteed, data sharing methods become unknown, and even future data value assessment and data product/commodity trading cannot be formed.

As pointed out in "Re-understanding the Library,"new library services should be: resource-based, technology-winged, demand-oriented, and service-king. This paper argues that for data diversity, environment is the foundation, technology is the wing, and standards are the essence. The combination of these three ensures the healthy development of data diversity and, on this basis, enables data application and data services, even intelligence services, think tank services, and intelligent services based on the integration of data, literature, and knowledge. Of course, "standards as the essence" here does not mean rigidly adhering to existing standards but rather that various principles and standards are the original foundation for safeguarding data diversity.

5 Application Representations of Data Diversity Theory

Data diversity involves multiple disciplinary fields, including Business Intelligence (BI), databases, networks and information communication, data publishing, strategic planning (data strategy), data models (or data modeling), data governance, data quality, data literacy, small data applications, and intelligent data.

5.1 Data Diversity in Data Strategy

Data diversity is reflected in all aspects. In pollution monitoring, data diversity helps enterprises comply with environmental regulations. Data scientists can capture environmental data from enterprise operations and analyze it together with other operational data to create actionable insights that provide competitive advantages and improve business efficiency. However, a truly diversified data-driven strategy goes beyond existing or easily collected data to discover new insights from data beyond what is immediately available from an organization' s main activities and operations. For example, in marketing, advertisers analyze how, when, and where their products are discussed, photographed, and posted on social media to better understand customers. In agriculture, farmers have become accustomed to using satellite and meteorological data to determine optimal timing and locations for crops.

5.2 Application of Data Diversity in Big Data Clustering

Computer scientists are also actively exploring, developing, and utilizing data diversity. For example, researchers from MIT' s Computer Science and Artificial Intelligence Laboratory and MIT's Laboratory for Information and Decision Systems have proposed a new diversity-based algorithm that ensures when sampling subsets from massive datasets, each subset retains the diversity characteristics of the complete set. This algorithm can be applied to various recommendation scenarios, such as books or movies, and can also be used in large-scale learning. The attribute of data diversity has also played a key role in many other application scenarios, such as gene network subsampling, document summarization, video summarization, content-driven search, recommendation systems, sensor placement, news headlines or search result prompts, image or photo scene clustering, citation chain research direction identification, and clustering of biological sequences or multimedia data.

5.3 Data Diversity in Small Data

Data diversity is also reflected in the small data domain. Although small data lacks a unified definition, diagnostic data and species research data all belong to small data. Therefore, some research data can also be considered a type of small data. In September 2021, the U.S. Cybersecurity and Emerging Technologies Agency released a research report titled "The Enormous Potential of Small Data AI," which pointed out that small data methods are AI approaches that require only small datasets for training, applicable in situations with limited data or no labeled data available, reducing dependence on collecting large real-world datasets. Small data methods include transfer learning, data labeling, artificial data generation, Bayesian methods, and reinforcement learning, which can be used in image recognition, machine learning, and other fields. Among these, transfer learning, data labeling, and active learning all conform to the characteristics of data diversity described earlier.

6.1 Pain Points for Researchers in the Open Science Data Environment

In the open science and open data environment, researchers face increasing pain points and difficulties. First, researchers and research teams need to address more and more data deposit tasks, including developing data management plans, opening data, submitting metadata, long-term preservation, as well as dealing with research integrity, research ethics, and performance evaluation.

Second, data disclosure has become a pressure for researchers. Studies on open data have found that research teams and their members also need to consider external funding and relevant norms when facing data disclosure. They require someone to provide full-process data consulting services rather than simple guidelines or best practices.

Finally, data reuse is difficult to achieve. The ideal state of data reuse or the ideal data ecosystem is that researchers can generate new data or databases after utilizing open data and share them with others. However, research has shown that some researchers become hesitant and reluctant when faced with data openness, making data reuse challenging.

6.2 Challenges for Data Librarians in the Open Science Data Environment

Librarians need to re-understand data diversity, identify researchers' data pain points, and help them solve the aforementioned problems and troubles, but they also face a series of challenges.

First, data librarians face challenges in data management capabilities, including storage, management, deposit, and preservation. Of course, data management capabilities require support from information infrastructure and business support from data librarians. At the same time, different stages of the research process require different data support.

Second, librarians face the challenge of communicating data ethics, including laws, regulations, policies, and agreements. As data librarians, they should understand the Copyright Law, Data Security Law, and Personal Information Protection Law; as well as data management measures, publishing management regulations, and electronic publishing management regulations; they should also understand relevant macro policies, such as the intellectual property powerhouse construction outline, opinions on the prosperity and development of academic journals, and the talent-strong country strategy; as well as Creative Commons (CC) licenses, free software licenses, and database usage agreements.

Finally, under the requirement of data as a production factor, data librarians may also need to understand data value-added businesses, including commissioning, exchange, trading, and negotiation. Although currently widely applied data mainly comes from the communications and e-commerce fields, it is believed that in the near future, scientific data will also be traded, and issues of data rights confirmation in research data will emerge.

6.3 Resource System Construction and Services Related to Library Development

As a place for collecting, circulating, and serving information resources, libraries inevitably need to break through the boundaries of traditional knowledge resources in an era where everything is data. Libraries have a long tradition in describing resources, providing access, building collections, and supporting long-term management of digital resources. Some libraries have also begun to participate in the entire lifecycle of data development, integration, and utilization, presenting and analyzing data within a broader mission and service scope. From the perspective of data governance, society directly faces data, and data

directly affects society, with libraries playing roles not as intermediaries but as drivers, facilitators, and assistants. Libraries can fully apply their accumulated experience in the literature domain to guide from technical, legal, and ethical rules. From the perspective of knowledge service library science theory, combining practical experience to demonstrate theory and requiring theoretical guidance for practice, data diversity is needed as a theoretical support for data services, because after literature services, information services, and intelligence services, data services are the final piece of the knowledge service puzzle. In the intelligence value chain of data-information-intelligence-decision-evaluation, data should be regarded as the starting point of intelligence work.

The current open science ecosystem has evolved from the first generation of literature repositories and data repositories for user storage, retrieval, and use, to the second generation of data products featuring citation links, metadata links, and third-party vocabulary links between literature and data. Through standard-based interoperability rather than metadata interoperability, the third-generation open science ecosystem is building an "organic growth body" of software, code, data, literature, citations, and evaluation content. Currently, the ultra-large metadata integration formed in the data and literature domains is developing toward similar data product directions. Prototype data products include: data, datasets, metadata, linked data, semantic data, open government data, research data, and data papers and publishing.

As a resource integration body, libraries should consider data diversity in data resource construction planning to enable data to be used by as many future users as possible and maximize data value. For existing data resources, data diversity should also be considered—that is, how data can be used in various scenarios as much as possible. Unlike literature, once data is built, it has its limitations or cannot be better used if data diversity is not considered during the planning stage of data resource construction. Conversely, if data resources are established under the guidance of data diversity principles and can be used by users to benefit research, a virtuous cycle of data resource development and utilization can be continuously carried out. Data diversity aims to enhance data' s discoverability, accessibility, interoperability, and reusability, which aligns with the connotation of the FAIR principles for research data. Libraries emphasizing data diversity can also ensure to a certain extent that research data follows FAIR principles; conversely, libraries starting from following FAIR principles can also ensure the use of common metadata systems or coding systems to describe, annotate, and archive research data, thereby enhancing data diversity.

For librarians or data librarians, engaging in collection knowledge database construction, management, and promotion policies has become a primary responsibility. In this process, exploring and developing metadata standards as best management practices, emphasizing data quality, accessibility, and interoperability is crucial. Librarians can attempt to recognize the importance of data diversity in data management plan practices and ensure research data diversity by improving the completeness of methods, policies, and standards in data

collection, description, organization, and storage. Additionally, librarians can provide embedded data support services to help researchers develop data plans, organize and process data, analyze and visualize data, and preserve data, providing seamless supporting services for data users and producers, which also enhances data diversity from an external manifestation perspective. Librarians have rich experience in information organization and can actively transfer this experience to the data domain (especially to research data or small data management), actively participating in and striving to perform well in data description, data labeling, or data cataloging to provide better data processing and management services for intelligent intelligence systems.

7 Conclusion

Diversity is of great significance and far-reaching impact. All forms of cultural diversity are competitive differentiation factors closely related to economic prosperity. Data diversity, as a form of cultural diversity, can only be truly recognized and realized in the data era, enabling organizations to better adapt to new ideas, new technologies, and new social and economic challenges. From the perspective of libraries and librarians, data diversity is the cornerstone of data services provided by libraries and librarians, the starting point for data intelligence work, and also the development opportunity for libraries and librarians to deeply participate in data-driven scientific discovery.

Note: Figure translations are in progress. See original paper for figures.