ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202307.00372

Post-print of Investigation and Analysis of Foreign Data Commons Platforms

Authors: Wu Yawei, Zhang Xiangxian

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] To investigate and analyze the data management models of foreign Data Commons, providing references for the construction of China's Data Commons. [Method/Process] By reviewing and synthesizing the development status of Data Commons domestically and internationally, comparing and analyzing the gaps between them, and taking the US INRG Data Commons as a case study, this paper examines its data space management model from the perspectives of principles and agreements, database and user interfaces, and data identification and linkage, thereby proposing strategies for the construction and development of China's Data Commons. [Results/Conclusion] Based on the case study and the current status of China's data sharing platforms, specific recommendations are proposed regarding overall planning, construction objectives, key issues to address, overall DC architecture, and user services.

Full Text

Preamble

Investigation and Analysis of Foreign Data Commons Platforms
Wu Yawei, Zhang Xiangxian
School of Management, Jilin University, Changchun 130022

Abstract

[Purpose/Significance] This paper investigates and analyzes the data management models of foreign Data Commons (DC) platforms to provide reference for building China's data commons. **[Method/Process]** By reviewing and summarizing the development status of data commons at home and abroad, comparing and analyzing the gaps between them, and taking the U.S. INRG Data Commons as an example, this paper examines its data space management

model from the perspectives of principles and protocols, database and user interface, and data identification and association, and proposes strategies for the construction and development of China's data commons. [**Result/Conclusion**] Combining the case study with the current status of China's data sharing platforms, specific recommendations are put forward regarding overall planning, construction goals, problems to be solved, DC overall architecture, and user services.

Keywords: Data Commons, Data Management, Data Services
Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2019.18.016

The powerful engine of the data era continues to drive science and technology and society forward, and the phrase "data is the new fuel" illustrates the important resource value of data [1]. Data is a valuable asset to be mined for individuals, organizations, and nations alike, and the most successful organizations in the future may be those that can leverage data resources to maximize tangible or intangible assets [2]. As data becomes increasingly valuable, data acquisition, analysis, sharing, and application for scientific research and decision-making have become extremely important and have received growing attention from countries and organizations. In recent years, China has also launched data management projects, such as the "National Scientific Data Sharing Project" initiated by the Ministry of Science and Technology in 2003 [3], which developed data sharing construction in different fields, and the "Outline for Promoting Big Data Development" issued by the State Council in 2015 [4], which proposed China's data sharing strategy. However, to date, teams and researchers in different fields still face problems in data acquisition, analysis, and sharing. In many research projects, issues related to privacy, property rights, and management in data sharing make the data acquisition and usage process cumbersome and lengthy, consuming significant time and effort from researchers and hindering the more comprehensive application of data to solve practical problems. Therefore, an efficient and sustainable data management model is needed to improve researchers' efficiency in acquiring, analyzing, and sharing data. As a large, interoperable data sharing platform, foreign Data Commons (DC) provides researchers and other users with a new research model that integrates heterogeneous data sources, analytical methods, and third-party applications, which can significantly enhance and expand the speed and scope of scientific data discovery for researchers or groups, thereby improving individual and team research productivity.

Thus, it is necessary for China to introduce foreign DC data management models and develop corresponding DC data services.

2. Connotation, Characteristics, and Theory of Data Commons

Foreign research and construction of DC originated around the 1970s, first applied in the medical field and later extended to other fields such as scientific research, economics, and social policy, primarily to address prominent data management issues in various domains. Over the years, theoretical research on DC has been continuously enriched and improved, covering DC planning, management, operation, development, and related regulations, which has enriched the connotation of DC while ensuring its distinctiveness.

2.1 Connotation of Data Commons

In the field of foreign data science, DC is a network infrastructure for locating, storing, and analyzing data, and more importantly, a data sharing space that facilitates research groups to use common methods and tools to analyze and share data [5]. Taking the Genomic Data Commons (GDC) in the medical field as an example [6], GDC is a data sharing platform that promotes precise analysis in oncology. It is not only a database or tool but also an extensible knowledge network for importing, standardizing, and optimizing the use of genomic and clinical data from various cancer research programs. Therefore, this paper argues that the connotation of DC is: an efficient data ecosystem platform that aims to serve science and users, strictly follows data regulations, solves data management problems for various data users, and integrates data acquisition, aggregation, identification, analysis, application, and sharing functions.

2.2 Characteristics of Data Commons

DC is widely applied in foreign data management fields and has demonstrated the following distinctive features through continuous development and practice:

Function and Role: DC emphasizes the integrated construction of functions such as data integration, association, discovery, review, analysis, application, and sharing, thereby creating a professional data ecosystem. It is committed to breaking data monopolies, enhancing researchers' reuse and innovation of data, and thus mining the deep value of data.

Data Processing Methods: When processing different data, DC follows a unified standardized process: (1) Pre-processing, including data entry, review, and screening, and standardizing and harmonizing data through metadata or data dictionaries; (2) Mid-term processing, including data analysis, visualization, transparency, and anonymization according to user needs; and (3) Post-processing, including data sharing, publishing, association, and other operations.

User Services: DC primarily provides services to users through physical spaces and virtual platforms. Physical spaces focus on deep engagement, communication, and mining user needs, while virtual platforms not only provide necessary

data management and analysis methods but also offer training and guidance for users to develop and research their own data analysis methods and tools. This improves data accessibility and interoperability while enhancing users' data literacy and awareness.

2.3 Progress in Data Commons Theory Research

2.3.1 Foreign Data Commons Theory Research Foreign theoretical research on DC started earlier and is relatively comprehensive, which can be summarized into three aspects:

(1) DC Design and Implementation. In recent years, new achievements have been made in DC design and implementation research, including DC framework design, solutions to funding issues, and applications in various fields. Regarding DC construction, F. Molinari et al. [7] and J. Mansell et al. [8] both designed a complete blueprint for DC and proposed that DC is a high-trust, low-cost data sharing platform that could replace other data management models in the future. Regarding funding issues, R. L. Grossman [9] provided guidance and recommendations on DC funding issues in the field of scientific research. For DC applications, S. P. French et al. [10] proposed creating DC for social big data, while the U.S. National Cancer Institute [11] and S. L. Volchenboum et al. [12] both supported the creation of DC for childhood disease data to enable in-depth cancer research.

(2) DC Management and Operation. In terms of DC management and operation, foreign scholars have analyzed current problems that need to be addressed. For example, S. A. Sansone et al. [13] proposed standardization and sharing principles for DC data; C. Bizer et al. [14] proposed solutions for DC metadata standardization, RDFa, and microformats; N. Purtova [15] clarified the sharing boundaries and social dilemmas of DC; other scholars have studied how to improve user interfaces and manage data, such as M. Morgan et al. [16] who proposed improving the interface for DC communication with users; Z. Su et al. [17] who made recommendations on how DC can improve disease data management; and C. Scott et al. [18] who proposed how to effectively analyze datasets in DC.

(3) DC-related Regulations. P. N. Halphin et al. [19] studied the regulations that DC must follow during operation and management, pointing out that DC must rely on relevant laws for sustainable development. OCLC [20] proposed that governments and relevant institutions should formulate policies to promote user acquisition, analysis, and sharing of DC data, encourage users to manage personal databases on DC, and provide user-attributed data based on regulations. J. Yakowitz [21] discussed the tragedy caused by obstacles to data sharing and proposed that data sharing should pay attention to privacy and security laws to minimize adverse consequences arising from privacy and property rights issues.

2.3.2 Domestic Data Commons Theory Research Domestic theoretical discussions on data sharing platforms can be summarized into two aspects:

(1) Introducing and Learning from Foreign Theoretical Achievements. The introduction and learning from foreign theoretical achievements can be categorized into: (a) Internal construction strategies, including platform construction and service models, functions and characteristics, goals and content, and personnel data literacy; and (b) External condition support, including policies and regulations, funding sources, institutional cooperation, and development and limitations. For example, Song Xiufen et al. [22] analyzed the characteristics, functions, and limitations of three famous university data platforms abroad and proposed that China should build from aspects such as platform functions, policy support, data standards, education and training, and cooperation and exchange. Qin Dan [23] and Wanyan Dengdeng [24] both analyzed data sharing platforms in British and American universities and provided development suggestions from aspects such as platform introduction, policy formulation, data service segmentation, funding sources, and construction models. Yang Helin [25] and Yin Shenqin [26] respectively introduced and evaluated foreign data platform models, ideas, characteristics, advanced functions, metadata standards, and online analysis functions.

(2) Research on Construction and Development of China's Data Commons. The content mainly includes two aspects: First, top-down analysis, which evaluates the functions, characteristics, and services of typical built data platforms, such as Zhu Ling et al. [27] who discussed the construction process of Peking University's open data platform, and Yin Shenqin et al. [28] who evaluated the system selection and functions of Fudan University's data platform. Second, bottom-up design, which provides suggestions from basic conditions such as platform construction systems, service systems, evaluation systems, data management, and policy formulation. For example, Deng Zhonghua et al. [29] designed a construction model for data sharing platforms under the "Internet+" environment from the perspectives of ensuring information security, expanding service content, and creating a sharing atmosphere; Liu Ziheng et al. [30] proposed building scientific research data sharing platforms based on subject service platforms or institutional repositories, formulating sharing policies, and improving data literacy; Liu Guifeng et al. [31] constructed a data platform evaluation index system from four aspects: platform construction foundation, data, management functions, and effects and impacts.

It can be seen that domestic and foreign research on data commons has different focuses, reflecting the gap between the two in theoretical research on data sharing platform design, implementation, management, and operation: (1) In design and implementation theory, foreign research focuses most on the preliminary design and preparation of a complete blueprint for data commons, particularly concerning funding issues, framework design, and clarifying what problems the data commons can solve later and how to solve them. In China, due to the late start in building and developing data commons, theoretical re-

search mainly explores implementation and operation frameworks suitable for its own conditions while learning from the construction experience of countries such as Britain and the United States, adopting a strategy of exploring, designing, and practicing simultaneously. (2) In management and operation theory, foreign research focuses first on analyzing and solving the dilemmas and problems of data commons, followed by managing and improving data and user services in the commons. For datasets, it requires following standardization and sharing principles, using metadata, RDFa, microformats, and other methods for data standardization to meet users' various data application needs, which belongs to the optimization stage. Current domestic research focuses on system selection, infrastructure construction, service improvement, evaluation system construction, policy formulation, and exploring specific paths for sharing space management and operation. Although some typical data spaces such as Fudan University and the Chinese Academy of Sciences have been put into operation, they are still in the construction and development stage.

3. Practical Development of Data Commons at Home and Abroad

3.1 Foreign Data Commons Practical Development

After decades of construction, development, and improvement, foreign DC has been applied in various fields and countries. The United States started construction earlier and has gradually spread to the United Kingdom, Australia, and other countries, forming DC management models with national characteristics. Through investigation, this paper summarizes the data management models of eight typical foreign DCs from the perspectives of research fields, functional characteristics, and operation models, as shown in Table 1 .

As shown in Table 1, the practical fields of DC cover basic science, healthcare, public services, and education. DC has made important contributions in discovering potential social problems, mining citizen needs, influencing public policy, and providing decision-making support. To varying degrees, DC also plays the role of a think tank, providing decision-making basis for researchers and governments through data management and analysis, enabling users to make more informed decisions.

(1) Function and Characteristics. DC has the ability to acquire, curate, apply, and share data for researchers. Under the premise of relevant data policies, it plays a role in serving society and the public. After years of development and improvement, the data management model of DC is also changing. It reduces the complexity and cost of data acquisition for users, improves data analysis quality, and simplifies data application processes, such as using graphical display technology to achieve visual analysis, link analysis, lineage analysis, and impact analysis of data, and realizing visual display of application integration relationships between systems to provide users with multi-level and fine-grained analysis results. In addition, DC breaks the limitations of data acquisition and

analysis, uses common infrastructure to analyze and share data, and provides an interoperable platform for scientific research groups. For example, a U.S. non-profit company [32] developed DC cloud computing infrastructure to support scientific research, such as open scientific data sharing clouds, with users including universities, non-profit organizations, companies, and government agencies.

(2) Data Management and Analysis. As a new tool, database extension, or type of network infrastructure, DC lowers user data acquisition costs and complexity, improves data analysis quality, and simplifies data application processes. These functions are all reflected in the user interface, allowing researchers to interoperate with these platforms, establish interfaces with DC through REST API programming to query and download data, drive current data portals, create index views of DC data models for projects, files, and cases, result analysis diagrams, sharing path diagrams, and collect relevant information. Therefore, DC particularly emphasizes the construction and improvement of user interfaces.

(3) Management and Operation Model. Most DCs strictly control their design, system selection, functional and service planning, and related rules during the initial construction phase, such as designing a complete implementation plan, formulating data management lifecycle processes, and related technical rules and protocols. In particular, through user incentive policies, they encourage various users to jointly participate in DC management and operation. Due to adequate preliminary preparation, the later-built website platforms, virtual spaces, and even physical companies often have more complete functions and better services.

3.2 China's Data Commons Practical Development

In the past decade or so, China has been building and developing data commons. Earlier examples include the Tsinghua University China Center for Economic Research, which began operations in 2009. In recent years, data spaces at Fudan University and Beihang University have begun to take shape. As shown in Table 2 , this paper selects eight representative data sharing platforms in China and summarizes their concepts and goals, functions and services, and construction and management mechanisms through investigation and analysis of data source coverage, platform functions and services, and cooperating institutions.

As shown in Table 2: (1) **Concepts and Goals.** The construction concept of China's data commons revolves around providing massive data resources, data services, decision support, and software development for their affiliated institutions or the public. The main goal is to achieve data storage and management, such as the software design and development services provided by the Huazhong University of Science and Technology Scientific Data Center. (2) **Construction Mechanism.** This can be roughly divided into two aspects: First, cooperative construction, which is adopted by most institutions, such as the Fudan University Social Science Data Research Center cooperating with Harvard University's

Dataverse to build a data center, and the National Earth System Science Data Sharing Platform. Second, independent construction based on the institution's own characteristics, such as the Beihang University Data Sharing Platform built through intra-institutional cooperation. (3) **Functions and Services.** In terms of functions, China's data commons mainly focus on data integration, storage, standardization, retrieval, analysis, and sharing, basically covering the main stages of the data lifecycle. In terms of user services, China's data commons mainly serve universities, research institutions, and government users. Some built platforms provide basic services such as user registration, login, access, download, and data analysis, while a few provide value-added services such as application, collaborative research, and decision support.

3.3 Comparative Analysis of Data Sharing Platforms at Home and Abroad

This paper selects five typical data commons from both China and abroad and compares them from three aspects: platform management and operation, platform functions, and user services, as shown in Table 3 . The comparison reveals that compared with foreign platforms, China's data commons construction still has gaps and deficiencies.

(1) Platform Management and Operation. Compared with foreign platforms, China's data platform management and operation are relatively weak, lacking a complete set of data governance planning, data lifecycle processes, reasonable management structures, and various protocol designs. For example, except for the DDI standard for metadata at the Fudan University Social Science Data Platform, other platforms have no clear specifications, which directly affects the practical application of data spaces, resulting in only a few functions being available to users and weak practicality. In terms of overall planning and policy protocols, although China has formulated macro-level policies and layouts for data management, it lacks meso- and micro-level management norms to guide and incentivize data management development, such as platform construction elements, usage and evaluation norms, personnel incentive policies, user norms, and necessary mandatory measures, thus slowing down the data management process and affecting implementation effectiveness.

(2) Platform Functions. Most domestic data sharing platforms have relatively complete functional settings for the early and late stages of data processing, mainly including data integration, storage, analysis, and sharing. However, in practical application and operation, these functions are not open to users, and some functions are still closed or need improvement. For example, most platform functions only include data retrieval (some with single retrieval methods and missing advanced search options), navigation (with some empty or broken links), download (only small amounts of data allowed for download, or requiring cumbersome application procedures), analysis and visualization (low openness of online data analysis and visualization leading to reduced user utilization), and sharing, which affect user experience and evaluation of the data platform.

(3) User Services. Most domestic data sharing platforms show a status of emphasizing functions over services in service planning and construction. Basic services such as data access, encouraging user participation, personalized settings, and user training are not fully realized and need improvement. Value-added services are even more lacking, such as collaborative research, connecting users and society, and decision consultation, which are fewer than abroad. Service methods are single and inefficient, with only a few providing decision support and user training services, such as the Fudan University Social Science Data Platform and the Huazhong University of Science and Technology Social Science Data Center. The reasons for these problems can be summarized as: lack of overall management and operation planning, insufficient user demand development, and lack of systematic guarantees such as concept promotion, demand research, and diversified service construction, leading to low user attention, lack of trust and interactivity, and thus making it difficult to develop comprehensive, thorough, and effective user-centered services.

4. Case Study of Foreign Data Commons—Taking the U.S. INRG Data Commons as an Example

4.1 Background of INRG-DC Launch

Worldwide, pediatric cancer, though rare, remains a difficult problem in the medical field. The lack of shared data on childhood cancer cases has created bottlenecks for in-depth research, and the emergence of DC has found a transformative approach to support this scientific research. In 2004, representatives from North American, European, and Australian/Japanese childhood cancer organizations collaborated to form the International Neuroblastoma Risk Group (INRG) to launch a childhood cancer data DC to find the best treatment methods through data analysis and sharing [41].

4.2 INRG-DC Management Model

The INRG assembled an expert team and, referring to the FAIR principles of scientific data management—Findability, Accessibility, Interoperability, and Reusability—formulated a set of data management models for INRG-DC. The main contents are as follows:

4.2.1 Overall Planning. INRG formulated a complete set of plans for designing, building, and managing DC, including: (1) Early-stage planning, covering funding, construction framework, management, and operation; (2) Mid-stage planning, including standardizing data lifecycle management processes and improving data management systems to ensure the scientific, systematic, hierarchical, and sustainable development of the data sharing space; and (3) Late-stage planning, mainly involving indicator-based evaluation of platform functions and user services as a basis for improvement.

4.2.2 Database and User Interface. Researchers at the University of

Chicago's Center for Research Informatics (CRI) designed and built a database for INRG phenotype data and opened a user front-end interface. Based on the protocol, anyone can query and use relevant data. Since 1980, the database has accumulated data from more than 18,000 patients and is regularly updated. In addition to basic data, the database can also filter biospecimen data through APIs to predict availability, greatly improving data acquisition speed and accuracy, simplifying the interaction process between the database and users, and enabling direct connection to target data.

4.2.3 Data Dictionary. Based on relevant standards and rules, INRG established a classification and analysis system and a data standardization system for pediatric cancer patient DC data. Together with statisticians from each region, they created a standard data dictionary, mapping all data elements into this framework, mainly standardizing and homogenizing clinical data from 8,800 patients diagnosed worldwide between 1990-2002, and enabling full utilization of the data to lay a foundation for data analysis and association.

4.2.4 Data Identification and Association. For data identifier issues, the Children's Oncology Group (COG) assigns a Universal Sample Identifier (USI) to each data point. The USI is associated with subsequently generated samples, data, and other information, enabling all samples in the target dataset to have USI links back to data in the INRG database, directly associating datasets and ensuring that various types of data can be directly invoked and associated with each other, allowing users to complete more comprehensive comparisons and analyses of data and broadening the scope of data usage.

4.2.5 Data Review and Curation. The DC operations center is a curation and approval process that requires users to formally access data through a portal. Data access is managed separately by the DC review mechanism, through which INRG clinical data and NCBI genomic data can be obtained. The DC then stores clinical data as an object, first reviewing data quality and quantity through a review model, then launching a virtual machine to use command-line tools for data analysis, preventing users from obtaining erroneous results due to using incorrect data [42].

4.2.6 Principles and Protocols. INRG convened relevant experts to formulate a set of operational principles and protocols for INRG-DC. Related protocols include: (1) Co-design rules, requiring DC to cooperate with technical experts, scientists, users, and relevant government organizations to exchange suggestions for DC implementation; (2) Metadata rules, requiring DC to have metadata, vocabulary control, and standardized database elements to make data easily searchable, discoverable, and associable; (3) Anonymization and security protocols, requiring DC to understand user purposes and needs, ensure appropriate and secure allocation of shared resources to users, and prevent data misuse through anonymization.

4.2.7 User Evaluation and Demand Analysis. DC completes multi-level analysis of data through internal and external data visualization analysis tools

and algorithms, such as visualization of data application frequency, related graphics visualization, and diagnostic testing, thereby analyzing internal and external characteristics of data at different levels. Users can also understand the data processing process through mind maps and even personally operate genomic data, such as extracting data from DC through genomic DC application programming interfaces, comparing it with clinical data, completing subsequent analysis and processing to generate results, and finally providing users with the most reasonable decisions and recommendations based on the results. This process can both explore and meet the needs of different users and serve as evaluation indicators obtained from users at each operational 环节.

4.3 INRG-DC Construction Achievements

4.3.1 User Data Ecosystem. INRG-DC has created an ontological data ecosystem that encourages data integration, analysis, sharing, and application. This data ecosystem attempts to achieve a complete data ecological cycle model of data integration, aggregation, analysis, and sharing based on maximizing stakeholder interests, gradually eliminating users' concerns about data property rights and privacy, establishing user trust in DC, and thus promoting the transformation from changing user awareness to promoting user behavior, thereby driving progress and development in data science.

4.3.2 Improved Data Lifecycle Process. DC has accelerated the powerful process of alliance-based collaborative discovery, data development, data attribution, data analysis, and data sharing, enhancing data interoperability and accessibility. The comprehensive collaborative development of DC infrastructure, policies, and processes has enabled the identification of children who need the most aggressive treatment while reducing the risk of ineffective treatment.

4.3.3 Creation of DC Management and Operation Environment. The development of genomic analysis and the democratization of data storage and computing resources have provided an ideal computing environment for DC to manage pediatric cancer data, enabling the collection, standardization, and aggregation of interconnections within DC of different phenotypic, genomic, and other data from pediatric patients. Today, the operation of DC in a sustainable environment not only promotes data application but also has a positive impact on pediatric cancer research, providing novel solutions for diagnosing and treating children with tumor diseases.

It is evident that INRG-DC plays an important role in helping solve data management problems in the medical field, thanks to its good construction, management, and operation. From overall DC planning to specific implementation, it has not only formed a series of data functions and service models covering database and user interface, data identification and association, data review and curation, and principles and protocols, but also achieved results including building a data ecosystem and improving data lifecycle processes. More specifically and microscopically, it has realized integrated management of plat-

form construction and user services. Therefore, when solving data management problems, China should appropriately learn from and introduce the construction and management model of INRG-DC to set overall plans, specific goals, and address main problems, building and formulating DC architecture and construction strategies adapted to China's data environment to provide users with more efficient and complete services.

5. Overall Framework for China's Data Commons Construction

5.1 Overall Planning, Goals, and Problems to be Solved by DC

5.1.1 Overall Planning of DC. As preliminary work for designing and building data sharing spaces, overall planning and layout are particularly important, such as formulating a complete set of preliminary design blueprints (including funding, construction framework, management, and operation) and complete data lifecycle processes, and improving data management systems to ensure the scientific, systematic, hierarchical, and sustainable development of data sharing spaces.

5.1.2 Goals of DC. The goals of China's DC construction mainly include: (1) Promoting the data sharing process—as a data sharing space, the purpose is to maximize data management and application to solve practical problems; (2) Optimizing user services—by providing user-centered data services to overcome data management obstacles users previously encountered; (3) Promoting communication and cooperation—by encouraging and facilitating communication and cooperation between data and users to fully discover and mine data value; and (4) Expanding data service fields—DC should act as a third party connecting data and users and expand service categories based on data regulations.

5.1.3 Problems to be Solved by DC. The construction and implementation of data sharing spaces aim to help researchers and other users solve problems related to integrating, standardizing, maximizing value, and optimizing allocation of large amounts of dispersed and potentially valuable multi-source heterogeneous data. Through effective DC management, target data from different fields can be quickly connected with data users, and advanced data analysis and management technologies can be used to achieve high integration of data, technology, and people. This enhances DC users' ability to process sensitive, large-scale, and unstructured data, building a complete and efficient data lifecycle model from data source determination, user-data interaction, data sharing and application, to reasonable distribution of stakeholder interests, enabling the most valuable data to be used by the most suitable users through DC as an effective channel, achieving the goals of data sharing and efficient application.

5.2 Overall Framework and Construction Strategy of DC

5.2.1 Overall Architecture of DC. China's data sharing space overall architecture (see Figure 1 [Figure 1: see original paper]) should consist of four parts: user layer, user interface layer, application and service layer, and data resource layer. First, various data users submit and share domain data resources to the internal DC through the user interface layer. Then, in the application and service layer, users manage target data, with the entire process completed under the curation of DC management and operation personnel to ensure balanced interests of the stakeholder layer. Finally, the mutual association, coordination, and interaction among all layers jointly constitute the complete management mechanism of the data sharing space platform.

(1) User Layer. This layer is the main object and population served by DC operations. According to the construction and application of DC in different fields and institutions (including universities, governments, enterprises, research institutions, etc.), users can be mainly divided into researchers, government personnel, university teachers and students, think tank personnel, enterprise personnel, and ordinary users. In addition, DC should clearly define and formulate interest distribution mechanisms and principles for various interest groups.

(2) User Interface Layer. The DC interface layer belongs to the basic operation layer of users, mainly responsible for directly interacting with users through interfaces such as websites and mobile terminals, processing user requests, providing various services to users, and coordinating with users to complete a series of activities such as data management and sharing. The layout of the user interface layer revolves around the user management system, mainly including login systems, permission control, personalized data push, and association with other databases. The user management system can also present a mapping relationship with the application and service layer and data resource layer and coordinate and feedback with each other. That is, through users' utilization of services such as data analysis, data peer-to-peer, and data sharing, the effects, performance of user interfaces, and quality of data resources can be fed back, and vice versa.

(3) Application and Service Layer. This layer is the core layer of DC, responsible for providing various data applications and services to users. It revolves around data ecosystem functions, mainly including applications and services such as data acquisition, data dictionary, data association, data analysis, data application, and data sharing. Addressing the shortcomings of domestic data management, this layer focuses on meso- and micro-level management, adopts a service-core strategy that integrates multiple applications and services to guide and incentivize data management development, serves users in broader fields, develops comprehensive, thorough, and effective data services, achieves precise identification and mining of users' actual and potential needs, cultivates users' data literacy and awareness in acquiring, analyzing, and applying data, and eliminates previous obstacles to users' acceptance of data services.

(4) Data Resource Layer. This refers to the data types and sources managed by DC and is the foundation of data management services. Data types include natural and social science data, such as economic, medical, and social data. Data sources include three categories: (1) Submission and sharing by different types of users; (2) Associated external data sources (such as data released by enterprises and governments); and (3) Data owned by DC itself. After determining data sources, data can be standardized and structured through DC data dictionaries, Digital Object Identifiers (DOI), or metadata to establish sustainable and analyzable data models that support peering and connection with other DCs or clouds, making data more easily usable, interoperable, network-capacity-level analyzable, and shareable.

5.2.2 Construction Strategy of DC

(1) DC Construction and Operation. This is mainly divided into construction entities and operation responsibility entities. The former refers to institutions or organizations that can undertake the task of building China's DC, such as libraries (mainly research and university libraries that expand traditional library services such as academic communication and consultation to the data stage, making library user services more targeted and practical [43]), information technology centers of various research institutions, and data management institutions. The latter includes software providers, government personnel, management teams, technical teams, various users, and other stakeholders. In addition, attention should be paid to balancing the distribution of interests among users and relevant interest groups.

(2) DC User Services. This is divided into three aspects: improving basic services, developing value-added services, and user incentive measures. Basic services include access permissions, encouraging user participation, personalized settings, and user training. Value-added services include collaborative research, connecting users and society, and decision consultation. User incentive measures include reducing data acquisition costs, push services (characteristic tools and services, etc.), promotion, and reward mechanisms. In addition, security restrictions should be considered to timely supplement gaps in DC space data, tools, and methods, and promote effective outreach and public participation in business.

(3) DC-related Protocols. DC responsible entities should formulate relevant rules for data space and user management with relevant institutions, such as: (1) Co-design rules: DC management needs to rely on collaborative management by technical experts, scientists, users, and relevant governments; (2) Metadata rules: DC needs complete and effective metadata, vocabularies, and data naming rules to make data searchable, discoverable, and associable; (3) Control protocols: manage users' data usage, and build Personal Information Management Systems (PIMS) to enhance user management mechanisms [44]; (4) Transparency and anonymization protocols: manage user needs transparently, allocate shared resources safely and reasonably for users, and anonymization can prevent data misuse.

References

[1] GROSSMAN R L, HEATH A, MURPHY M, et al. A case for data commons: towards data science as a service[J]. Computing in science & engineering, 2016, 18(5): 10-20.

[2] Data Commons Launch Project [EB/OL]. [2019-01-09]. http://www.bio-itworld.com/2017/11/07/nih-launches-data-commons-pilot-with-9-projects.

[3] Zhang Xian'en. National Scientific Data Sharing Project[J]. Scientific Chinese, 2004(9): 11.

[4] State Council. Notice of the State Council on Issuing the Outline for Promoting Big Data Development [EB/OL]. [2019-01-09]. http://www.gov.cn/zhengce/content/2015-09/05/content_{10137}.htm.

[5] National Institutes of Health. Newly launched genomic data commons to facilitate data and clinical information sharing [EB/OL]. [2019-01-09]. http://www.nih.gov/news-events/news-releases/newly-launched-genomic-data-commons-facilitate-data-clinical-information-sharing.

[6] Wu Yawei, Wei Lai. Development and preliminary construction of foreign Data Commons[J]. Information and Documentation Services, 2017, 38(6): 41-48.

[7] MOLINARI F, MORELLI N, TORNOTOFT L K, et al. OpenDataLabs: new infrastructures for open data commons [EB/OL]. [2019-01-09]. https://www.forskningsdatabasen.dk/en/catalog/2372370890.

[8] New Zealand Project. Data commons blueprint [EB/OL]. [2019-01-09]. http://datacommons.org.nz.

[9] GROSSMAN R L. Data Commons Guidelines [EB/OL]. [2019-01-09]. https://www.healthra.org/wp-content/uploads/2018/08/Data-Commons-Guidelines_{Grossman}8_{2017}.pdf.

[10] FRENCH S P, BARCHERS C V. Designing a data commons for urban big data [EB/OL]. [2019-01-09]. https://www.rd-alliance.org/final-report-income-streams-data-repositories.html.

[11] VOLCHENBOUM S L, HAWKINS D, FRAZIER L, et al. Building pediatric cancer data commons [EB/OL]. [2019-01-09]. https://ascopubs.org/doi/full/10.1200/EDBK_{175029}.

[12] VOLCHENBOUM S L, COX S M, HEATH A, et al. Data commons to support pediatric cancer research[J]. American Society of Clinical Oncology Educational Book, 2017, 37(24): 746-752.

[13] SANSONE S A, MCQUILTON P, ROCCA-SERRA P, et al. FAIRsharing_{{{working}}*{{with}}*{{and}}*{{for}}*{{the}}*{{community}}*{{to}}*{{describe}}*{{and}}*{{link}*
[EB/OL]. [2019-01-09]. https://www.researchgate.net/publication/326462185.

[14] BIZER C, MEUSEL R, PRIMPELI A. The web data commons microdata, RDFa and microformat dataset series [EB/OL]. [2019-01-09]. https://link.springer.com/chapter/10.1007/978-3-319-11964-9_{91}.

[15] PURTOVA N. Health data for common good: defining the boundaries and social dilemmas of data commons [EB/OL]. [2019-01-09]. http://link.springer.com/chapter/10.1007/978-3-319-11964-9_{91}.

[16] MORGAN M, DAVIS S R. Genomic data commons: a bioconductor interface to the NCI genomic data commons [EB/OL]. [2019-01-09]. https://github.com/seandavi/GenomicDataCommons.

[17] SU Z, BERTAGNOLLI M M, SARTOR A O, et al. A novel, open-access method to improve disease management in patients (pts) with Merkel cell carcinoma (MCC) [J]. Journal of clinical oncology, 2018, 36(15): 215-255.

[18] SCOTT C, WALTERS S, EDDIE S, et al. VDJServer: a cloud-based analysis portal and data commons for immune repertoire sequences and rearrangements [J]. Frontiers in immunology, 2018, 9(39): 976-1002.

[19] HALPHIN P N, READ A J, BEST B D, et al. OBIS-SEAMAP: developing a biogeographic research data commons for the ecological studies of marine mammals, seabirds, and sea turtles [J]. Marine ecology progress series, 2006, 77(316): 239-246.

[20] EVANS B J. Barbarians at the gate: consumer-driven health data commons and the transformation of citizen science [J]. American journal of law & medicine, 2016, 42(4): 651-685.

[21] BAMBAUER J Y. Tragedy of the data commons [J]. SSRN electronic journal, 2011, 62(2): 120-135.

[22] Song Xiufen, Deng Zhonghua. Research on institutional repositories based on data curation[J]. Library Science Research, 2016, 31(2): 44-48.

[23] Qin Dan. Investigation and analysis of social science data management and sharing service platforms in the UK and US[J]. Library and Information Service, 2014, 58(8): 67-75.

[24] Wanyan Dengdeng. Research on scientific data management and sharing policies in Australian universities[J]. Journal of Information Resources Management, 2016(1): 30-37.

[25] Yang Helin. New ideas on institutional repository construction in American university libraries from the perspective of data curation—Insights from DataStar[J]. Journal of Academic Libraries, 2012(2): 23-28, 41.

[26] Yin Shenqin, Zhang Jilong, Zhang Ying, et al. Research on system selection for social science data management and service platforms—Taking Fudan University Social Science Data Platform as an example[J]. Library and Information Service, 2013, 57(19): 92-96.

[27] Zhu Ling, Nie Hua, Cui Haiyuan, et al. Construction of Peking University Open Research Data Platform: Exploration and practice[J]. Library and Information Service, 2016, 60(4): 44-51.

[28] Yin Shenqin, Zhang Jilong, Zhang Ying, et al. Research on system selection for social science data management and service platforms—Taking Fudan University Social Science Data Platform as an example[J]. Library and Information Service, 2013, 57(19): 92-96.

[29] Deng Zhonghua, Huang Yating. Research on the development of China's scientific data sharing platform under the "Internet+" environment[J]. Information Studies: Theory & Application, 2017, 40(2): 128-132.

[30] Liu Ziheng, Zeng Liying. Investigation and comparative analysis of scientific research data management and sharing platforms in Chinese universities[J]. Information and Documentation Services, 2017(6): 90-95.

[31] Liu Guifeng, Zhang Yu, Liu Qiong. Construction of an evaluation index system for open scientific research data platforms and case studies[J]. Library and Information Knowledge, 2019(1): 21-31.

[32] U.S. Open Data Cloud Alliance [EB/OL]. [2019-03-09]. www.open-science-datacloud.org.

[33] Fudan University Data Center [EB/OL]. [2019-03-09]. https://dvn.fudan.edu.cn/home/static/profile.jsp.

[34] Beihang University Data Sharing Platform [EB/OL]. [2019-03-09]. http://etc.xzit.edu.cn/01/19/c56a281/page.htm.

[35] Tsinghua University Data Sharing Platform [EB/OL]. [2019-03-09]. http://www.chinaz.com/news/2016/0105/492077.shtml.

[36] Chinese Academy of Sciences Computer Network Information Center [EB/OL]. [2019-03-09]. http://www.nsdata.cn/resource/list?code=1803710.

[37] Chinese Academy of Sciences Data Sharing Platform [EB/OL]. [2019-03-09]. http://www.geodata.cn/.

[38] Wuhan University Library Data Sharing Center [EB/OL]. [2019-03-09]. http://www.lib.whu.edu.cn/kxsj/aboutus.htm.

[39] Huazhong University of Science and Technology Scientific Data Center [EB/OL]. [2019-03-09]. https://cmis.csdc.info/toabout.action.

[40] Tsinghua University Economic and Social Data Center [EB/OL]. [2019-03-09]. http://www.sem.tsinghua.edu.cn/sercent/jjshsjzx.htm.

[41] International Neuroblastoma Risk Group. INRG data commons [EB/OL]. [2019-03-09]. http://europepmc.org/abstract/MED/28561664.

[42] HEATH A P, GREENWAY M, POWELL R, et al. Bionimbus: a cloud for managing, analyzing and sharing large genomics datasets [J]. Journal of the American Medical Informatics Association, 2014, 21(6): 969-975.

[43] Wei Lai, Gao Xiran. Role positioning of university data librarians under the background of big data[J]. Information and Documentation Services, 2015(5): 90-94.

[44] MANSELL J, LAKING R, MATHESON B, et al. Data commons blueprint: a high-trust, lower-cost alternative to enable data integration and reuse [EB/OL]. [2019-03-09]. http://datacommons.org.nz, 2017.

Author Contributions:
Wu Yawei: Topic selection, writing, and revision;
Zhang Xiangxian: Paper review, revision, and final approval.

Journal of Library and Information Science 2019 Topic Guide

The *Journal of Library and Information Science* is an authoritative academic journal with over 60 years of history in library and information science and related fields, dedicated to publishing and exchanging theoretical, technical, methodological, and application innovation achievements in library science, information science, and related interdisciplinary fields. We welcome submissions of innovative research papers with theoretical contributions or application value that are thoughtful, creative, methodological, and empirical.

2019 topics include but are not limited to:
1. Research on the development of China's library and information cause in the 70 years since the founding of the People's Republic of China
2. Functions and influences of library societies (associations) in library development
3. Pre-research on China's library and information cause "14th Five-Year Plan"
4. Role positioning of libraries in the OpenScience era
5. Functions and characteristics of library scientific communication in the new media era
6. Role of libraries in reconstructing academic communication systems
7. Artificial intelligence and smart libraries/smart services
8. Research on library laws and related laws in China and abroad
9. Theory and practice of library embedded services
10. From information literacy education to innovation literacy education
11. Resource organization and services across LAM (libraries, archives, museums) fields
12. Impact and effectiveness evaluation of new library construction and space reconstruction
13. Research and practice of library scientific and technological achievement transformation
14. Key issues in next-generation institutional repository construction
15. Characteristics and requirements of library data resource construction
16. Data-driven new-generation library system construction
17. Innovation and application of information science theory and methods
18. Reform of intelligence systems under the overall national security concept

19. Innovation in intelligence analysis theory and methods
20. Intelligence service capabilities from a big data perspective
21. Library and information science and think tank construction/services
22. Think tank services and decision-making consultation service capability building
23. Theory and method system of computational intelligence science
24. Technology and methods of data management and services
25. Data governance and national intelligence security strategy
26. Intelligence sharing mechanisms in military-civilian integration
27. Micro-mechanisms and macro-phenomena of information behavior
28. Regional and industrial intelligence service models and mechanisms
29. Utilization and value assessment of multi-source information resources
30. Theory and practice of Altmetrics

Journal of Library and Information Science Magazine
December 2018

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

Source: ChinaXiv — Machine translation. Verify with original.