Research on Metadata Creation Services for Comprehensive Scientific Data Repositories (Postprint)

Authors: Huang Guobin, Wang Tao

Date: 2023-04-01T16:02:58+00:00

Abstract

[Purpose/Significance] The lack of knowledge regarding scientific data metadata and efficient, user-friendly metadata creation services impedes researchers' sharing and reuse of scientific data. Comprehensive scientific data repositories, characterized by large storage volumes and broad user bases, provide metadata creation services that offer valuable insights for ameliorating this predicament. [Method/Process] This study investigates and analyzes the metadata creation services of six comprehensive scientific data repositories recommended by Springer Nature and Scientific Data, examining both service content composition and implementation models to summarize their characteristics and best practices. [Results/Conclusion] The metadata creation services offered by comprehensive scientific data repositories exhibit five major characteristics: preserving tradition while fostering innovation, striving for simplicity while highlighting distinctive features, emphasizing metadata knowledge dissemination and capability transformation, fully ensuring data democracy, and focusing on related resource organization while encouraging data citation. Their service models balance usability and usefulness with metadata knowledge dissemination, offering significant implications and reference value for the development of data repositories and metadata creation services in Chinese library and information institutions.

Full Text

Research on Metadata Creation Services in Generalist Research Data Repositories

Authors and Affiliation

Huang Guobin and **Wang Tao**
School of Government, Beijing Normal University, Beijing 100875

Abstract

[**Purpose/Significance**] The lack of scientific data metadata knowledge and efficient, user-friendly metadata creation services hinders researchers from sharing and reusing scientific data. Generalist research data repositories, due to their large data storage volumes and broad user base, offer metadata creation services that provide valuable reference for addressing these challenges. [**Method/Process**] This study examines six generalist research data repositories recommended by Springer Nature and Scientific Data, analyzing their metadata creation services from two perspectives: service content composition and implementation model, and summarizes their service characteristics and best practices. [**Result/Conclusion**] The metadata creation services provided by generalist research data repositories exhibit five key characteristics: inheriting tradition while innovating, pursuing simplicity while highlighting distinctive features, emphasizing metadata knowledge popularization and capability transformation, fully ensuring data democracy, and focusing on related resource organization while encouraging data citation. Their service models balance usability, usefulness, and metadata knowledge dissemination, offering important insights and references for data repository construction and metadata service development in Chinese library and information institutions.

Keywords: research data; metadata creation services; generalist repository
Classification Number: G239.2
DOI: 10.13266/j.issn.0252-3116.2021.21.020

1. Introduction

Effective research data management requires robust technical support but depends even more on high-quality data organization, which in turn relies on the completeness and systematic nature of metadata describing the data. Dataset publication is the process of storing datasets and their associated attributes to make them accessible to user communities. As the starting point of dataset publication, researchers use metadata creation services within platforms to actively submit metadata required for publication. The quality and level of these services not only affect researchers' willingness to publish and share their datasets but also directly determine the metadata quality of published datasets. Metadata provides knowledge about what data exists, where it is located, and where it can be shared.

T. Carol and her team surveyed over 1,000 researchers on how they manage their research data in 2011 and 2015. The 2011 results showed that more than 50% of researchers had never used any metadata standards, and only 26% were satisfied with the scientific data metadata creation tools they used. The 2015 survey still

found that 47.9% of researchers had never used any metadata standards. The *State of Open Data Report 2019* indicated that 54.33% of researchers worldwide had never heard of the FAIR (Findability, Accessibility, Interoperability, and Reusability) principles, and 48% were unclear about how to apply data licensing agreements. Foreign research shows that researchers currently lack scientific data metadata knowledge and are not adept at creating metadata for their data, while metadata creation services and tools have shortcomings. These factors directly or indirectly affect the sharing and reuse of scientific data.

As the foundation for data browsing, retrieval, and sharing services provided by data repositories, how to provide quality metadata creation services is one of the core issues that must be considered in scientific data repository construction.

---

2. Literature Review

Using title searches in CNKI with the query "(Title:(metadata + scientific data)) AND (Title:service)", we retrieved 89 relevant documents. Through a top-down, layer-by-layer narrowing approach, we found that domestic research exhibits the following characteristics: (1) Many studies focus on scientific data services in foreign university libraries, emphasizing overall analysis and introduction of advanced foreign experiences. For example, Xiao Xiao et al. summarized five ways foreign libraries participate in scientific data services in the E-Science environment, while Wang Cuiping et al. identified seven major scientific data service projects in university libraries using five US, UK, and Australian universities as examples. (2) Early attention was paid to potential classification, indexing, and literature linking issues in scientific data description and organization. Scholars investigated and summarized technical measures and solutions explored by foreign stakeholders, such as Qian Peng et al., who proposed six key issues to be addressed in scientific data organization and services, and Qiu Chunyan, who summarized approaches and key implementation methods for linking journal literature with scientific data. (3) A few scholars focused on metadata services and metadata creation services provided by foreign university libraries. For example, Huang Xin et al. surveyed and analyzed scientific data metadata services in eight university libraries ranked in the top 100 by US News and summarized four forms of metadata creation services. (4) Some scholars studied metadata schemes in typical data repositories but did not examine how metadata standards or schemes are applied in specific metadata creation services, such as Hu Fang's research on metadata schemes in four typical foreign data repositories.

Overall, domestic research on scientific data metadata creation services is limited. Existing studies mainly introduce, analyze, and import advanced models and experiences of foreign university library scientific data services at the macro level. The few studies on metadata creation services are confined to university libraries and lack systematic analysis and experience summarization of metadata creation services on external platforms. This study selects six generalist

research data repositories, focusing on their metadata creation services, and analyzes and summarizes their characteristics in combination with scientific data organization and literature linking issues proposed by relevant scholars. The aim is to provide insights and references for data repository construction and metadata creation service design and implementation in Chinese university libraries, ultimately enabling them to provide researchers with easy-to-use and high-quality scientific data metadata annotation solutions.

---

3. Research Objects and Approach

Scientific data repositories (RDR), also known as data repositories, centers, or platforms, serve as infrastructure for research data management and are generally divided into disciplinary and generalist types. According to Springer Nature's research data submission requirements for academic journals, disciplinary repositories are those where paper-associated data is submitted to academically recognized repositories based on the specific discipline of the paper, while generalist repositories are alternative options when no suitable disciplinary repository is available. In the context of interdisciplinary integration, generalist data repositories serve a broader community, typically providing scientific data creation, submission, storage, publication, and management services to the entire scientific community. They support storage of any type and discipline of scientific data, thus requiring higher construction standards and fully reflecting the construction level of a country's scientific data sharing infrastructure.

This study selected six repositories from the list of generalist research data repositories recommended by Springer Nature and Scientific Data as investigation objects (data collected until January 2021): Dryad Digital Repository (DDR), Figshare, Harvard Dataverse (HD), Zenodo, Mendeley Data (MD), and Science Data Bank (SDB). These repositories were selected because: (1) Their metadata creation services have clear boundaries with other services and are independent; (2) Their operators are diverse, not limited to university libraries; (3) They are mature platforms serving global researchers with large data volumes, indicating their service quality and capability are recognized by the academic community, making their metadata creation services valuable references; (4) SDB became the only repository on this list independently developed and built by China in September 2020, making the selected repositories representative both domestically and internationally. Using web-based investigation combined with literature review, this paper analyzes the metadata creation services of these six platforms from two aspects: service content composition and implementation model.

---

4. Analysis of Metadata Creation Services

4.1 Metadata Elements and Value Settings

The setting of metadata elements in platforms is primarily related to the metadata framework referenced during repository construction. When developing metadata schemes, repositories typically refer to one or more metadata standards to form their own schemes, expanding or reducing elements and values as needed. Zenodo and DDR explicitly state that their metadata element settings are based on the DataCite schema. HD primarily adopts and references three metadata frameworks: DDI (Data Documentation Initiative), DC (Dublin Core), and DataCite schema. Although the remaining three platforms do not explicitly state the metadata frameworks their element settings are based on, their specific metadata element settings show that they fully absorb and reference relevant metadata standards in natural science and social science fields while striving for simplicity.

The metadata element settings of the six platforms are basically the same [Figure 1: see original paper]. Elements can be generally divided into mandatory and optional elements. Mandatory elements are required for dataset submission and publication, while optional elements are additional attributes users attach to datasets beyond publication requirements. Both mandatory and optional elements can be further divided into compound and simple elements. Compound elements contain at least two sub-elements, such as the mandatory element "Author," which typically includes sub-elements like name, affiliation, ORCID, and email. The optional element "Related Works" usually contains three sub-elements: related resource identifier or URL link, relationship, and resource type. Both compound and simple elements can be further divided into repeatable and non-repeatable categories. Repeatable elements can be recorded multiple times. In all six platforms, the "Author" element is a repeatable compound element. Under compound elements, sub-elements can also be divided into mandatory and optional regardless of repeatability.

Value settings must ensure completeness of dataset description while considering user usability and value control. The metadata element values in the six platforms can be divided into three categories: (1) Fixed selective values provided through dropdown lists for direct selection; (2) External reference values that provide external reference standards for selection and recording; (3) User-recorded values with no reference information or only recording requirements, where users directly record based on existing knowledge or after reading element explanations and guidance documents. The first two categories reference controlled vocabularies or domain ontologies and absorb metadata attribute values from different terminology systems, providing normative references and tangible constraints for dataset description, while the third category fully embodies the principles of openness and freedom.

4.2 Metadata Element Organization Methods

Metadata element organization is the process of reorganizing and reordering originally loose and unordered metadata elements and values according to different functions and attributes. As the pivot in transforming user-created initial metadata into the platform's final standard dataset metadata, it has two major functions: (1) User guidance, enabling users to quickly understand the logical framework of scientific data metadata creation and integration within the platform; (2) Mediation and transformation, forming a user metadata creation template through the metadata element organization framework, which ultimately transforms user-created initial metadata into the platform's published standard dataset metadata.

After fully investigating typical dataset metadata annotation and organization schemes in natural and social science fields and selecting domain standard cases for comparison , we found that although their annotation and organization emphases differ due to disciplinary differences, their overall metadata element setting and organization frameworks are basically the same and show a trend toward standardization. This is achieved through standardized modular metadata organization methods that enable systematic integration of metadata elements and internal-external resource linkage.

Generalist data repositories do not completely copy the standard modular metadata element organization methods from natural and social science fields. Instead, they adjust, select, and innovate based on their own metadata element quantity, complexity, and platform functional positioning . Platforms mainly adopt three metadata organization methods: (1) Direct listing for platforms with fewer elements to demonstrate simplicity; (2) Modular organization for platforms with a moderate number of elements to reflect logical organization; (3) For platforms with many and complex elements like HD, an innovative method combining external layering and internal modular organization is used to integrate logical and hierarchical organization and achieve comprehensive description from macro dataset level to micro file level for citation purposes. Regardless of the method used, the ultimate goal is user usability, enabling users to efficiently and normatively complete metadata recording along a clear main line when creating metadata for scientific data.

4.3 Characteristics Analysis of Elements and Values

4.3.1 Mandatory Elements Serve Data Citation The design of metadata elements and values primarily aims to achieve complete description of dataset background information, personalized declaration of dataset rights, and comprehensive association between datasets and related resources. Metadata created by researchers for datasets provides critical information for data reuse and reproduction, determines when data is available and how to use it correctly, and is also a prerequisite for data citation. A comprehensive analysis of metadata elements and values across the six platforms reveals four main characteristics .

Overall, the mandatory elements required for dataset publication are relatively few, except for SDB, which currently has fewer collected datasets and thus requires relatively more mandatory elements to ensure data discovery. According to the *Joint Declaration of Data Citation Principles* released in 2014, a complete data citation format should include author, year, dataset title, globally unique identifier, repository name, and version, while the citation format encouraged by DataCite includes author, publication year, title, publisher, resource type, and identifier. The mandatory element sets required by the six platforms basically include all elements necessary for dataset citation, with excluded elements typically automatically generated upon data publication, such as DOI and publication time.

4.3.2 Setting Dedicated Elements to Ensure Data Reuse To fully ensure data reproduction and reuse, all platforms set separate elements or require users to provide information about data collection steps and methods, instruments and equipment used, etc., through element combinations in specific values. According to the specificity of statements, they can be divided into two categories: (1) Brief statements requiring users to briefly describe the above information under a single element value; (2) Detailed statements combining main element values for brief statements with sub-element values for detailed statements. For example, DDR uses the "Methods" and "Usage Notes" elements to fully ensure complete presentation of data background information, while MD, in addition to the mandatory "Description" element, specifically sets the sub-element "Steps to Reproduce" requiring users to detail their data collection process. The first category applies when platforms have large data volumes or are in initial construction stages, while the second category mainly applies to natural science datasets.

4.3.3 Element Combination Mechanism for Data Democracy Control In Wiley's survey, two important reasons researchers were unwilling to share their scientific data were: (1) Fear of negative consequences such as data misuse, legal or commercial repercussions; (2) Lack of recognition mechanisms for their work. The first reason arises because researchers lose control over their datasets after publication, while the second stems from insufficient incentives for data sharing and publication. To address these issues, the six platforms mainly adopt three solutions for flexible data control: (1) Mandatory licensing, such as DDR requiring all published datasets to use CC0 licenses because their datasets are mostly associated with journal papers and primarily serve peer review; (2) Multiple licensing options allowing users to flexibly select the most applicable license based on dataset type, such as MD dividing licenses into three categories: pure data, hardware, and software licenses, then allowing users to decide how datasets can be reused under specific licenses, including attribution, non-commercial use, share-alike, no derivatives, and combinations thereof; (3) Combined use of licenses with other methods, allowing users to set data protection periods or specific time nodes for public release, such as SDB. Figshare

further allows users to select partial or overall protection through "Embargo Type"—if users choose file-only protection, data files remain private during the protection period but metadata records are publicly accessible; if overall content protection is selected, both datasets and metadata records are private during the protection period. HD allows specific restrictions from dataset level to file level through the "Terms" compound element, and users can use the built-in tool Dataset Guestbook to collect key information like name, email, institution, and geographic location of downloaders when datasets are downloaded.

4.3.4 Diversified Data Association Methods All platforms emphasize describing relationships between datasets and other related resources. According to resource types, associations can be divided into three categories: single resource association, partial resource association, and multiple resource association, primarily achieved through URLs and DOIs to enable precise association between datasets and multiple types of resources. Considering the diversity of relationship types—including citation, reference, compilation, derivation, part, replacement, continuation, and description—MD and Zenodo attempt semantic description of relationships and achieve this through the combination of three sub-elements: related resource type + related resource identifier + relationship. Additionally, Zenodo and HD set up dedicated units to provide applicable specialized metadata elements for special associated publications such as conference papers, theses, books, or specific chapters to help users locate specific associated content.

---

5. Implementation Models of Metadata Creation Services

Currently, the main forms of scientific data metadata creation services domestically and internationally include six types: publishing guidance documents, embedding into research processes, self-submission forms, providing metadata file templates, providing software tools, and intelligent parsing of metadata configuration files. According to the degree of manual participation, they can be divided into three categories: fully manual creation, semi-manual creation, and automatic creation. The trend in scientific data repository metadata creation services is to integrate these forms according to the platform's own resource positioning and user needs to achieve automated and intelligent metadata creation.

5.1 Self-Submission Forms

Form content mainly includes metadata element names, identifiers, and explanations. Based on the completeness of element explanations covering elements, they can be divided into complete explanation type and partial explanation type. Platforms simplify service processes through concise forms, using clear symbol identifiers to grade element importance and supplementing with specific

element explanations to help users quickly enter and adapt to metadata creation environments without systematically learning professional metadata knowledge.

All six platforms mainly provide explanations including definition of meaning, normative recording examples, and filling suggestions. Unlike traditional academic resources, to systematically and comprehensively describe scientific data provenance information and increase reproducibility and usability, dataset descriptions typically include creation and transformation history metadata attributes to record how they were created and subsequent conversion and processing information. Both partial and complete explanation types provide brief, accessible explanations for elements related to this special attribute to enhance user understanding and guide users to correctly, normatively, and completely record relevant values .

Specifically, partial explanation platforms use heuristic questioning and direct statements to explain and illustrate relevant elements, mainly reflecting simplicity and fully considering readability while grading element importance. Complete explanation platforms use interactive explanations to explain and illustrate elements, mainly reflecting comprehensiveness and achieving immediate interactive answers to user confusion. For example, HD first provides clear definitions in interactive explanations for relevant personnel making different contributions in the dataset lifecycle, such as Author, Producer, Contributor, Distributor, and Depositor, and gives specific conceptual definitions and boundary divisions for time elements like Production Date, Distribution Date, Deposit Date, and Time Period Covered to ultimately eliminate ambiguity in element meanings.

5.2 Guidance Documents

Form content provides users with essential basic knowledge about metadata elements, mostly at the operational level. Platforms provide single or multiple guidance documents to meet users' needs for further understanding platform metadata information and solve practical problems encountered during metadata creation. Guidance documents are generally set up under FAQ documents, help documents, and data management and archiving documents, covering the role and significance of metadata, general content that should be included in metadata, standards and guidance schemes adopted by metadata, supplementary explanations of each element's meaning, and terminology systems used for normative control of relevant elements. Guidance documents aim to explain what scientific data metadata is and what metadata should be created for scientific data, emphasizing the popularization of scientific data and metadata background knowledge to enhance users' metadata creation capabilities.

---

6. Characteristics of Metadata Creation Services

Through systematic analysis of the service content composition and implementation models of the six generalist research data repositories, although their

services differ at the micro level, they overall exhibit the following characteristics:

6.1 Fully Absorbing and Learning from Domain Knowledge, Inheriting Tradition While Innovating

Generalist research data repositories fully learn from and absorb metadata standards in natural and social science fields when setting metadata elements and values. Based on their own functional positioning, element scale, and disciplinary focus, they transform and innovatively express traditional modular metadata organization methods from natural and social science fields in their metadata organization models.

6.2 User-Centered, Pursuing Simplicity in Service Content and Models

Generalist research data repositories abandon the complexity of specific metadata annotation schemes in natural and social science fields. Their service content strives to balance simplicity and metadata quality. Considering users with different knowledge backgrounds, their service models aim to balance simplicity, usability, and quality in metadata creation services. Through dual simplification of content and model, they achieve immediate publication of scientific data.

6.3 Emphasizing Scientific Data Metadata Knowledge Popularization, Valuing Knowledge Transformation and Capability Enhancement

These platforms locally penetrate scientific data metadata element knowledge in metadata creation practice, systematically introduce scientific data and metadata background knowledge in guidance documents, achieve penetration and popularization of partial and overall knowledge, and ultimately complete two-way flow and mutual promotion between metadata knowledge and metadata creation capabilities.

6.4 Focusing on Protecting Data Creators' Rights and Interests, Fully Reflecting Data Democracy

Through mandatory licensing and multiple licensing options, platforms provide policy-level support for reasonable data use. Beyond basic licensing for data democracy control, combined with user-defined terms, protection period settings, and built-in development tools, they improve and innovate to further strengthen data creators' rights protection and reduce data misuse risks.

6.5 Emphasizing Organization of Related Resources and Encouraging Data Citation

V. Timothy et al. studied over 500,000 papers published in PLOS and BMC and found that researchers storing their research data in repositories and linking it

with relevant papers increased average citation rates by about 20%. The *State of Open Data Report 2019* indicated that complete citation of research papers remains the strongest motivation for researchers to share their scientific data. All platforms provide dedicated metadata elements or modules to basically achieve association between datasets and multiple resources, with innovations in association methods and expressions beyond basic linking. After dataset metadata creation is completed, the system automatically generates multiple data citation formats, creating favorable citation conditions to enhance the impact of datasets and related academic achievements.

---

7. Implications for Data Repository Construction and Metadata Creation Service Development in Chinese Library and Information Institutions

The metadata creation services of generalist research data repositories demonstrate numerous advantages and characteristics in service content and implementation models, providing advanced domestic and international experiences for scientific data repository construction and metadata creation service development in Chinese library and information institutions. The following three aspects can be learned and innovated upon:

7.1 Based on Platform Positioning, Pursue Simplicity and Distinctive Features in Service Content

When constructing scientific data repositories, Chinese library and information institutions should selectively choose, absorb, and reference domain metadata standards according to their own disciplinary positioning and primary service targets. After fully investigating current difficulties and needs researchers face in dataset metadata annotation, they should develop distinctive metadata annotation and organization solutions while pursuing simplicity. Repositories primarily serving natural science researchers for journal paper-associated dataset storage and review can reference the metadata organization schemes of DDR and SDB. Those mainly serving social science researchers for dataset submission and storage can reference HD's scheme combining modular and layered metadata organization. Multi-disciplinary data repositories pursuing minimalist models with good publisher partnerships can reference the metadata organization schemes of MD and Figshare.

7.2 Fully Leverage Information Literacy Training Advantages of Library and Information Institutions to Achieve Service Model Breakthroughs and Innovation

Due to large data volumes and diverse dataset types, generalist research data repositories strive to balance service simplicity and usability. However, limited personnel, funding, and management resources mean they only provide user

training and guidance through guidance documents, which is clearly insufficient. Chinese library and information institutions can fully leverage their rich experience in information literacy education to conduct scientific data literacy training and guidance. To this end, they can build learning centers on institutional or platform websites: (1) Set up guidance document sections to systematically organize and summarize scientific data and metadata knowledge, categorize historical problems in user metadata creation, update existing problems timely, anticipate future possible problems, and provide detailed and feasible solutions; (2) Set up skills training sections to enhance user data skills, regularly release online and offline training notifications, and organize training resources such as videos, webinars, and written guides to show users how to search, understand, and correctly use datasets through diverse training forms and resource types, ultimately improving user metadata creation capabilities; (3) Set up technical resources sections to detail currently most commonly used software and tools in scientific data analysis, processing, collection, and metadata creation, and publish tool usage guides and links to help users understand and start using them to improve metadata creation efficiency and quality; (4) Attempt to set up scientific data management sections. As a core component of Data Management Plans (DMP), metadata schemes have been required by an increasing number of research funding agencies since NSF proposed DMP in 2011. China's *Management Measures for Scientific Data* also clearly requires science and technology plan management departments at all levels to establish special mechanisms for project acceptance. In the future, Chinese library and information institutions can integrate scientific data metadata creation services, DMP writing services, and data repository submission and storage services to fully leverage the role of scientific data metadata creation services in the research lifecycle. Metadata creation services can provide necessary metadata for DMP during early data collection phases, revise and supplement metadata during data collection, and form systematic and complete metadata required for data submission and storage after project completion.

7.3 Leverage Professional Knowledge Advantages of Library and Information Institution Staff to Innovate in Metadata Quality Control

Currently, metadata creation in generalist research data repositories mainly relies on manual input, supplemented by semi-automated selection and input of element values. Facing massive scientific data, they have not yet formed effective metadata quality control mechanisms. In 2019, NIH's Office of Data Science Strategy (ODSS) collaborated with Figshare on a one-year project to determine how biomedical researchers use Figshare to share and reuse NIH-funded scientific data. The project found that datasets reviewed by professional data librarians had 2.5 times higher download and view volumes than unreviewed datasets, indicating the importance of strengthening metadata quality review. When developing scientific data metadata creation services, Chinese library and information institutions can select appropriate metadata quality control mechanisms based on platform user and dataset submission scales: smaller-scale

repositories can allocate corresponding proportions of data librarians for manual review; larger-scale repositories can adopt automatic control as the main approach supplemented by manual review, developing corresponding software tools for preliminary review of dataset metadata format and content, and configuring intelligent metadata assessment systems for preliminary quality evaluation, with data librarians conducting secondary review of problematic dataset metadata from preliminary review and providing specific supplementary instructions and improvement suggestions.

References

[1] FERGUSON L. How and why researchers share data (and why they don't) [EB/OL]. [2021-03-20]. https://doi.org/10.6084/m9.figshare.3468365.v1.

[2] TENOUPIR C, ALLARD S, DOUGLASS K, et al. Data sharing by scientists: practices and perceptions [J]. PloS one, 2011, 6(6): e21101.

[3] TENOUPIR C, DALTON E D, ALLARD S, et al. Changes in data sharing and data reuse practices and perceptions among scientists worldwide [J]. PloS one, 2015, 10(8): e0134826.

[4] Digital Science. The state of open data report 2019 [EB/OL]. [2021-04-27]. https://figshare.com/articles/report/The_{{{State}}*{{of}}*{{{Open}}*{{Data}}*{{{Report}}*{{2019}}*/99

[5] Xiao Xiao, Lv Junsheng. Research on foreign library scientific data services in the E-Science environment [J]. Library and Information Service, 2014(2): 63-66.

[6] Wang Cuiping, Li Jialu. Current status and enlightenment of scientific data services in foreign university libraries [J]. Library Work and Study, 2017(10): 31-36.

[7] Qian Peng, Zheng Jianming. Preliminary exploration of scientific data organization and services in universities [J]. Information Theory and Practice, 2011, 34(2): 27-29.

[8] Qiu Chunyan. Research on linking services between journal literature and scientific data [J]. Information and Documentation Services, 2012, 56(17): 53-58, 114.

[9] Huang Xin, Deng Zhonghua. Research on metadata services for scientific data in foreign university libraries [J]. Library and Information, 2017(2): 84-90.

[10] Hu Fang. Metadata schemes in typical foreign scientific data repositories and their enlightenment [J]. Library and Information, 2015(1): 117-121.

[11] Springer Nature. Research data policies [EB/OL]. [2021-01-21]. https://www.springernature.com/gp/authors/research-data-policy/recommended-repositories/.

[12] Springer Nature. Generalist repositories [EB/OL]. [2020-12-09]. https://www.springernature.com/gp/authors/research-data-policy/repositories-general/.

[13] Computer Network Center, Chinese Academy of Sciences. ScienceDB becomes the only Chinese-developed repository recommended by Springer Nature [EB/OL]. [2021-01-26]. https://www.cas.cn/yx/202010/t20201010_{4762415}.shtml.

[14] DataCite Metadata Working Group. DataCite metadata schema documentation for the publication and citation of research data and other research outputs (Version 4.4) [EB/OL]. [2021-04-16]. https://doi.org/10.14454/3w3z-sa82.

[15] Yang Bo, Hu Liyun. DDI: a metadata standard for social science information organization [J]. New Technology of Library and Information Service, 2005(8): 7-11.

[16] UK Data Service. Quarterly labour force survey, April-June, 2021 [EB/OL]. [2021-09-03]. https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=8826#!/details.

[17] LIANG D J, ZHUANG M H, HU C Y, et al. China's greenhouse gas emissions for cropping systems from 1978-2016 [EB/OL]. [2021-09-04]. https://doi.org/10.1038/s41597-021-00960-5.

[18] Data Citation Synthesis Group. Joint declaration of data citation principles [EB/OL]. [2021-02-26]. https://www.force11.org/group/joint-declaration-data-citation-principles-final.

[19] Wanyan Dengdeng. Investigation and enlightenment of foreign scientific data repository metadata practices [J]. New Century Library, 2016(5): 81-84.

[20] SINGH G, BHARATHI S, CHERVENAK A, et al. A metadata catalog service for data-intensive applications [EB/OL]. [2021-09-06]. https://dl.acm.org/doi/10.1145/1048935.1050184.

[21] ROBINSON N, JIMENEZ E, TORRES D. Analyzing data citation practices using the data citation index [J]. Journal of the Association for Information Science and Technology, 2016, 67(12): 2964-2975.

[22] Dryad. Frequently asked questions [EB/OL]. [2021-09-06]. https://datadryad.org/stash/faq.

[23] Figshare. Tutorials [EB/OL]. [2021-09-05]. https://help.figshare.com/section/tutorials.

[24] Harvard Dataverse. Dataset and file management [EB/OL]. [2021-09-05]. https://guides.dataverse.org/en/5.6/user/dataset-management.html.

[25] Zenodo. Frequently asked questions [EB/OL]. [2021-09-05]. https://help.zenodo.org/.

[26] Mendeley Data. How can we help you? [EB/OL]. [2021-09-06]. https://data.mendeley.com/faq.

[27] Science Data Bank. Frequently asked questions [EB/OL]. [2021-09-06]. https://www.scidb.cn/en/faq.

[28] VINES T H, ANDREW R L, BOCK D G, et al. Mandated data archiving greatly improves access to research data [EB/OL]. [2021-04-25]. https://doi.org/10.1096/fj.12-218164.

[29] NIH. Data management guidance for CISE proposals and awards [EB/OL]. [2021-02-02]. https://www.nsf.gov/cise/cise_{dmp}.jsp.

[30] General Office of the State Council. Notice on issuing the Management Measures for Scientific Data [EB/OL]. [2021-04-20]. http://www.gov.cn/zhengce/content_{2018}-04/02/content_{5279272}.htm.

[31] ANAV G. NIH figshare instance highlighted use cases [EB/OL]. [2021-09-07]. https://figshare.com/articles/online_{resource}/NIH_{{{Figshare}}{{Instance}}}{{{Highlighted}}{{Us

[32] Bureau of Science Communication, Chinese Academy of Sciences. National standard "Data Paper Publication Metadata" project launch meeting held [EB/OL]. [2021-04-09]. http://www.bsc.cas.cn/sjdt/202103/t20210324_{4782140}.html.

English Abstract

Research on the Metadata Creation Service of the Generalist Research Data Repository

Huang Guobin, Wang Tao
School of Government, Beijing Normal University, Beijing 100875

Abstract: [Purpose/significance] The lack of research data metadata knowledge and efficient, easy-to-use metadata creation services hinders researchers from sharing and reusing research data. Due to the large amount of data and wide user orientation of the generalist research data repository, the metadata creation service provided by it has reference significance for improving the above dilemmas. [Method/process] Taking six generalist research data repositories recommended by Springer Nature and Scientific Data as samples, this paper investigates and analyzes their metadata creation services from two aspects: service content composition and service implementation model, and summarizes their service characteristics and advanced experiences. [Result/conclusion] The metadata creation services provided by the generalist research data repository have five characteristics: following tradition and having some innovations, striving for simplicity and highlighting its own characteristics, paying attention to the popularization of metadata knowledge and ability transformation, fully ensuring data democracy, paying attention to the organization of related resources and encouraging data citation. Its service model not only pays attention to the ease and usefulness of service but also takes into account the popularization of metadata knowledge, which has important enlightenment and reference for the construction of data repositories and the development of metadata creation services for library and information institutions in China.

Keywords: research data; metadata creation services; generalist repository