Journal of Electronics and Information Science (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2022.070307
ISSN 2371-9524 Vol. 7 Num. 3

Metadata-Based Management and Application Research of Biological Information Resource

Yan Yang[1,*], Li Wang[1], Bo Yao[1], Mingnuan Han[1], Rui Liu[2], Xiaoling Yue[2], Wei Yu[2], Yi Li[3]

[1]Yunnan Provincial Academy of Science and Technology, Kunming, China
[2]Yunnan Landian Technology Co., Ltd, Kunming, China
[3]Yunnan University of Finance and Economics, Kunming, China
**Corresponding author*

Keywords: Metadata, scientific data, biological information resources, information resource management

Abstract: Chinese biological resource ranks top in the world both in biomass and biodiversity. With the rapid increasing storage capacity of biological information resource in Internet era, it's necessary to discuss how to form a scientific and effective management, how to spread data application and how to add value of biological information resource. In this paper, we first explained the concept of biological information resource's metadata. Then, we discussed the significant problems of metadata's construction such as system's dispersion, standard's disunity, security's loss and the unstable quality. At last, we gave some suggestions to promote the Chinese biological resource's development according the successful development of the others.

1. Introduction

The rapid development of digital technology and computer science has provided digital and electronic management for data management. How to promote the open sharing, reuse and value realization of big data in various industries and fields has become the focus of digital strategies in various countries, while opening up new management paths for scientific data. China promulgated *the Action Outline for Promoting the Development of Big Data* in 2015 , which clearly proposes to develop scientific big data, gradually realize the open sharing of the scientific data acquired or generated by public welfare scientific research activities supported by the national finance, construct scientific big data national major infrastructure, and realize the authoritative collection, long-term preservation, integrated management and comprehensive sharing of national important scientific and technological data. Since then, the State Council and the Chinese Academy of Sciences have successively issued the *Measures for the Management of Scientific Data* (GB F [2018] No. 17) and the *Measures for the Management and Open Sharing of Scientific Data of the Chinese Academy of Sciences (for trial implementation)* (KF B Zi [2019] No. 11) to continuously expand the innovative applications of scientific data into various industries and fields.

Biological resources are an important material basis for human survival and development, mainly including plants, animals, microbial organisms and their constituent communities, populations and ecosystems[1], all countries in the world as a community of life attach great importance to the conservation and research of biological resources. China's long-term research on biodiversity has accumulated a large amount of basic information on plants and animals, biological specimens, genetic data, etc., which are stored in catalog systems of different fields and departments. With the promotion of the construction of informationization of biological resources, the application of metadata technology for recording, management as well as sharing is of great significance to realize the scientific values, economic values and social values of biological information resources.

Based on the concept of metadata, this paper focuses on identifying the main problems of biological information resources in China at the data level, analyzing the main problems and obstacles in the open sharing of biological resources at the present stage, referring to the advanced experience in the application of metadata management, and proposing corresponding countermeasures for promoting the construction of metadata for biological information resources.

2. Biological Information Resources Metadata

2.1. Metadata

The concept of metadata can be traced back to the Directory Interchange Format (DIF) published by NASA in 1987, which is considered to be an important element of data resource directory exchange, and then began to be widely used in the International Directory Network (IDN) as a way to record and exchange scientific data. In 1995, the first conference on metadata was held in Dublin, Ireland, which promoted the academic community to conduct relevant research on metadata from different perspectives and fields. Metadata is considered as data about that data or data that describes other data[2], as an important carrier to record the relevant characteristics and attributes of data resources, and is an important basis for effective organization and application of data. In terms of content, a complete metadata is composed of datasets and resources; in terms of structure, metadata can be divided into structured data and unstructured data[3]. Structured metadata is an important basis on which computers of different systems can recognize, exchange, integrate, and process data.

The earliest research results on metadata in China were published in 1994[4], and in the following years, the fields of computer, library intelligence and geoscience focused on the localized application of foreign research results, and paid attention to the compatibility, interoperability and storage retrieval of different metadata. With the development of big data and cloud computing technologies, data has become a fundamental and strategic resource, and the management and utilization of metadata has become a common concern in various fields. In this context, scholars start from the concept of metadata and conduct research around the standard specification, the quality evaluation, and the description methods of metadata, etc.[5]; some scholars also based on the perspective of resource utilization and study the role and problems in data standards, data opening, and data sharing of metadata in the fields of educational information resources, geospatial resources, and governmental information resources, etc.[6-9]; other scholars discuss metadata management systems, data format standards, technical systems, etc. from a technical perspective[10,11]. It can be seen that with the development of the times, the concept of metadata has gone beyond the traditional instrumentalist scope of describing and managing resources, and expanded to become a management method or standard for information resources.

In the field of scientific data, metadata plays an important role in the process of describing records, long-term preservation, open access and sharing[12]. The National Science Foundation (NSF) of the United States Federal Government applied metadata management to develop the Data Management Plan and imposed requirements on its councils that all scientific data generated by projects must include every detail (Table 1) [13].

Table 1: U.S. Scientific Data Management Plan

Data Management Plan	Data Type	Sample data, physical specimens, software, course materials and other materials that will be generated during the project
	Data Standards	Include metadata standards and content standards
	Data Acquisition and Sharing Policy	Privacy protection, confidentiality, security, intellectual property rights and other rights claims
	Data Archiving and Preservation Plan	Project application without data management plan will not be accepted

Scientific data refers to the original basic data reflecting the essence, characteristics, and change law, etc. of the objective world obtained in scientific and technological activities (experiments, observations, probes, surveys, etc.) or by other methods, as well as various data sets that are systematically processed and organized according to the needs of different scientific and technological activities. However, regardless of the application scenario or the research approach, the essence of metadata is still the description of data resources.

2.2. Biological Information Resources Metadata

Biological resources are the most basic material basis for human reproduction and development, mainly including animals, plants, microbial organisms and their constituent communities, populations and ecosystems. For biological information resources, its sources mainly include two kinds: one is the scientific research data generated through experiments, observations and investigations during the implementation of national science and technology plan programs and various scientific research and practice by scientific workers; the other is the business data collected and managed by government departments for a long time. Various departments at all levels, mainly scientific research institutions, have collected and stored a huge amount of biological data, formed a large-scale biological information resource system, managed it effectively and promoted its extensive sharing to maximize its value and make it become an urgent problem in this field. At the same time, facing the fact that the multi-source, networked and dynamic characteristics of resources are increasingly prominent, metadata technology and directory management have become the crucial modern scientific data management methods.

At present, China has constructed the "China Biobank", "Chinese Academy of Sciences Strategic Biological Resources Programme (BRP)" and other open platforms of biological data, bringing together 72 biological repositories of 40 research institutions of the Chinese Academy of Sciences, with more than 7.43 million biological resources data[17]. However, due to the short time of domestic metadata application in the field of open biological data and the lack of unified specification, there are still some biological open data with poor specification, single format and difficulties in utilization and so on. The relevant management method can refer to the US NSF regulations for scientific data management of the Biotechnology Industry Organization(BIO): each data management plan must cover the following issues: what types of data will be collected, the standards to be adopted, and the duration of data preservation; what physical facilities or network resources (including third-party resources) will be used to preserve the data; metadata formats, carriers, and dissemination methods, etc. of providing data sharing.

The biological information resource metadata is a structured description of biological information resource composition, characteristics, standards, storage and other information, which aims to let users understand the data attributes they need to query or use as soon as possible. Referring to the existing metadata classification methods, the metadata of biological information resources can be divided into four categories: management, description, preservation and technology[14] (Table 2), such metadata classification methods are easy to apply to the construction and management of biological information resources.

Table 2: Classification of Biological Information Resources Metadata

Classification	Metadata Elements	Functional Purpose
Management	Inputer	Provide basic information for data resource management and maintenance
	Affiliation	
	Contact Information	
	Input Time	
	Update Time, etc.	
Description	Character Characteristics	Provide data features, attributes and other information is easy for user identification and use
	Living Environment	
	Functional Purpose	
	Cultivation Management, etc.	
Preservation	Comments	The directory index information of data is easy for the user to find the data storage location quickly.
	Index	
	Data Cataloging, etc.	
Technology	Preservation Format	Data used to develop and manage data on a daily basis
	Safety Verification	
	Access Rights	
	Data Backup, etc.	

Among them, management metadata mainly provide basic information for data resource management and maintenance, such as the data inputer, the affiliation of the data inputer, contact information, input time, last update time, etc.; description metadata are mainly used to provide information such as characteristics and attributes of data that are easy for users to identify and use, such as trait characteristics, growth environment, medicinal value, etc.; Preservation metadata is the index information of the data, which is easy for users to find the data storage location quickly, such as data cataloging, index, annotation, etc.; technology metadata refers to the technical data of preservation format, system, security verification, etc.[15]

3. Problems in the Management and Application of Metadata of Biological Information Resources

With the rapid development and popularization of digital information technology, a large amount of biological information resources has been collected, aggregated, and organized by various departments at all levels, especially by scientific research institutions, and biological data has become a huge "data mine" for scientific research and industrial innovation development. The management and application of biological information resource metadata is an important basis for the realization of biological data value. Joanne Evans and Barbara Reed have identified the ability to establish a sustainable framework for creating and managing records metadata as one of the key challenges for recordkeeping in digital and networked environments.[16] However, the construction process of the existing biological information resources is affected by the scattered system construction and the lack of creation standards, updating mechanisms, access rights, security assurance, quality standards and evaluation systems of metadata, which both hinder the opening and sharing of biological data and restrict the development of biological industry.

3.1. The Construction of Biological Information Resource System is Fragmented

Influenced by the traditional institutional functions of detached functions, government departments or research institutions construct information systems based on the work and research needs of their own units or departments, and due to the lack of a global view at the early stage of construction, various types of biological data are constructed under different directory systems as independent "islands", and a large amount of biological data are scattered in the hands of government departments, research institutions and even researchers at all levels, and the fragmented distribution of biological information resources affects the sharing, application and innovation of resources by government, enterprises and society.

3.2. Lack of Data Creation Standards and Update Mechanisms

The current innovation and update of biological information resources data are mainly departmental, and at the level of data creation, there is a lack of macroscopic principles and ideas of metadata creation due to the limitation of traditional system construction and usage scope, which leads to diverse data creation formats and single application scenarios. In the context of promoting the sharing and use of biological information resources, the original data creation concept and creation standards are unable to meet the application requirements. In terms of data update, the data update mechanism is not sound, resulting in a large number of "zombie data" and invalid data, which affects and restricts the innovative application of data in various industries.

3.3. Lack of Data Access Rights and Security Assurance

In some frontier research fields, the scope of sharing and use of biological data is limited, and there is no clear data access authority and access scope for who can access the data or what data are conditionally shared. There are clear specifications and requirements on how data can be used safely after acquisition. A large number of data collection units or collectors are reluctant to share or dare not share because they are worried about data security, resulting in a large amount of valuable data that cannot be shared and opened, which restricts research and development and innovation in the field of biological resources and limits the realization of the value of a large amount of data.

3.4. Lack of Data Quality Standards and Evaluation System

The quality standards of metadata directly affect the effectiveness of managing and utilizing data. Due to the business differences among units and research departments, as well as the influence of informatization conditions and levels, the quality requirement standards in the metadata of the information resource management system have different standards, large differences, and semantic ambiguities in description information, resulting in the inability to realize effective data docking in the process of data sharing. The low interoperability results the need to spend a lot of time on data cleaning and alignment, and most organizations can only re-collect data according to their own standards, resulting the problems of repeated collection and multi-standard collection. On the other hand, there is a lack of effective evaluation index system for metadata quality: it is impossible to provide quality standards in the development of construction process, which makes it difficult to improve the quality of data construction, thus restricting the application of data.

4. The Main Path to Promote the Management and Application of Metadata of Biological Information Resources

4.1. Clarify the Access Rights to Biological Information Data

Firstly, clarify the restrictions on accessing or downloading data. It is important to make users of data clarify whether they can access or download data, and explain why. For example, according to the *General Data Protection Regulation* (GDPR) established by the EU, the UK government proposes that data involving personal privacy should not be made publicly available without the conditions or scope of institutional permission, and all data providers should specify the Use constraints of data. If there are no restrictions on the data, the publisher of the data resource should also state that it has taken into account the tips on data security. Secondly, data can be taken in various licensing options for use. For example, add one or more use department licenses below the data, indicate that each data can optionally add a different license use restriction element, and clearly indicate the circumstances in which each option applies.

4.2. Improve the Quality Level of Metadata of Biological Information Resources

Metadata management cannot be effective if the quality of metadata is not taken into account.[17] Data quality is a quantitative description of the integrity, uniqueness, consistency, timeliness, accuracy and validity for data[18]. Firstly, high-quality metadata is a guide for users to retrieve biological data resources, which can quickly help users to find the required data and understand the relevant information of the dataset. For example, a certain type of animal or plant metadata should at least contain information about who generated the data, who maintains it, and how it is used, as well as hint that the data is part of a series and provide more relevant information about the series, and use parent IDs or links to link the data to other datasets in the series to other datasets whose data belong to a series so that they can quickly find other dataset, which will reduce their efforts to find other data resources that may be useful for their needs. Secondly, metadata is the basic description of the biological data resource, what the data resource contains, where it comes from, how often it is updated, the licensing requirements for data use, and the data quality standards, and users can access this information without accessing the resource itself, allowing users of the data to more quickly determine whether the data resource is suitable for the intended use.

4.3. User-oriented Metadata Creation and Update

The creation and updating of metadata is the key to ensure the effective use of data. The construction of metadata for biological information resources should have five characteristics of being innovative, concise and distinctive, emphasizing metadata knowledge popularization and capability transformation, fully guaranteeing data democracy, paying attention to the organization of related resources and encouraging data citation based on the essence of metadata, and its service model should pay attention to both the ease of use and usefulness of information resources and metadata knowledge popularization[19].

When creating metadata, firstly, it should be created with the principle that metadata can be discovered and used to the maximum extent. By establishing a standard specification for data creation, on the one hand, it provides guidance to data publishers on how to create metadata; on the other hand, it allows helping data users to discover data and evaluate its usefulness to them. Knowing how the data is created will help them use it and prevent many users from wasting time to get data that they cannot use. Secondly, the principle of extensive use and reuse of data should be used to pay attention to the use cases of current users and understand the use needs of potential users to promote data use and data value. Thirdly, it is important to establish different update mechanisms for different types and needs of data, maintain and update metadata dynamically throughout the life cycle, provide standards and methods for data collection, processing, modification and update, and understand whether data quality meets users' usage needs for continuous improvement.

4.4. Construct a Data Genealogy of Biological Information Resources

Using a detailed genealogy to explain the generation of a certain type of data and the stages it has gone through, that is, the full life-cycle genealogy of biological data. On the one hand, confirming the information of data sources and integrating a large amount of data from different fields or institutions, the information from multiple sources can not only help in using the data but also play an important role in tracing the data; on the other hand, tracing the update, dissemination and use path of information based on blockchain and other information technology, dynamically understanding the panoramic genealogy of data from creation to value realization, can provide macro and panoramic guidance for decision making, regulation, services and innovation.

4.5. Multi-party Joint Application of New Technologies to Broaden the Scope of Data

Draw on Biosphere Reserve, rhoen (Germany) of United Nations Educational, Scientific and Cultural Organization, take a holistic view of biological resource protection and data management, strengthen cooperation and communication among departments and units, break down "administrative barriers", and enhance business synergy; consider the collection, database and analysis of monitoring data and metadata by means of metadata, eco-regionalization and geostatistics, etc., and combine multi-technical tools to break down "technical barriers"[20]; improve the interoperability of data by adopting a unified and standardized metadata standard system, and solve the "information barriers" and the phenomenon of estrangement that difficult to use data together, difficult to exchange information, difficult to share information, and difficult to collaborate on information processing. And construct an open data ecosystem, carry out metadata governance, guarantee metadata quality and solve interoperability problems ultimately.

5. Conclusion

As an important part of scientific data, the biological information data is a national strategic basic resource, and the effective management and application of biological information resources metadata can effectively support the biological industry, scientific research and further development. At present, from the concept of metadata, the concept and connotation of biological information resources metadata are identified and discussed in the context of the problems in the construction of biological information resources metadata in China, and the corresponding construction paths are proposed to provide useful help and inspiration for promoting the construction of biological information resources and the sharing and reuse of data.

References

[1] Shi Xiaoliang, Chen Ke, Lu Chenxi, (2018) He Dan. Review of Biodiversity Value Evaluation Research. China Forestry Economics, 6, 104-108.

[2] Zhao Qingfeng, Ju Yingjie. (2003) A Review of Domestic Metadata Research. Modern Information, 11, 42-45.

[3] Qi Tianjiao, Feng Huiling. (2021) The Connotation, Characteristics and Principle Analysis of Semantic Organization in the Process of Archival Datalization. Library and Information Service, 65, 9, 3-15.

[4] Qin Xiao. (1994) Meta-data Dictionary and its Implementation. Chinese Journal OF Computers, 2, 81-87.

[5] Liu Zhifeng, Wang Jimin, Li Qian. (2022) A Review of Metadata Quality Evaluation Research. Information studies: Theory and Application, 45, 7, 7.

[6] Jin Jiaqin. (2005) Educational Metadata and its Application in Online Education. Journal of Intelligence, 1, 60-61.

[7] Sun Guangzhi. (2001) Application of Metadata on Sharing Information Resources. Information Science, 7, 763-764+779.

[8] Shen Tiyan, Cheng Chengqi. (2000) Research on Environmental Metadata Standards and Environmental Information Sharing Model in China. Environmental Protection, 5, 32-34.

[9] Zhai Jun, Zhai Wei, Pei Xintong, Li Jianfeng. (2021) Construction of Metadata Standards for Sharing and Opening Government Data in U.K. and the Enlightenment. Journal of Intelligence, 40, 4, 132-138+186.

[10] Wang Yuwa, Si Li. (2021) Research on the Construction of the Metadata Standard System of Multilingual and Shared Economic Management Database for Countries Along the Belt and Road. Research on Library Science, 3, 44-53.DOI:10.15941/j.cnki.issn1001-0424.2021.03.006.

[11] Qian Yi. (2012) On the management strategy of metadata scheme of electronic document center. Archives Science Bulletin, 6, 76-79. DOI:10.16113/j.cnki.daxtx. 2012.06.022.

[12] Si Li, Zeng Yueliang. (2017) Research Progress on Institutional Research Data Repositories Abroad. Journal of the China Society for Scientific and Technical Information, 36, 8, 859-870.

[13] Si Li, Xing Wenming. (2013) Scientific Data Management and Sharing Policies in Foreign Countries: Investigation and Inspiration to Us. Information and Documentation Services, 1, 61-66

[14] Tao Shuilong, Wang Zhen, Tian Lei, Bai Wei, Ren Wenge. (2016) Classification and Scheme Design for Electronic Records and Archives Metadata. Archives Science Study, 6, 83-90.

[15] Xu Yonglong, Long Jiaqing. (2020) Study on the optimization strategy of China's electronic document long-term preservation format standard--and a comparative analysis with Britain, the United States, Canada and Australia. E-government, 8, 113-124.

[16] Joanne, Evans, Barbara, et al. (2008) Interoperable data. Records Management Journal, 18, 2, 115-129.

[17] Elouataoui, W., ElAlaoui, I., Gahi, Y. (2022). Metadata Quality in the Era of Big Data and Unstructured Content. In: Maleh, Y., Alazab, M., Gherabi, N., Tawalbeh, L., AbdEl-Latif, A.A.(eds) Advances in Information, Communication and Cyber security. ICI2C2021. Lecture Notes in Networks and Systems, vol357. Springer, Cham.

[18] Sheng Xiaoping, Tian Jing, Xiang Guilin. (2020) Research on Data Quality Governance in Open Sharing of Scientific Data. Library and Information Service, 64, 22, 11-24.

[19] Huang Guobin, Wang Tao. (2021) Research on the Metadata Creation Service of the Generalist Research Data Repository .Library and Information Service, 65, 21, 131-140.

[20] Schroeder W, Pesch R, Schmidt G. (2006) Identifying and closing gaps in environmental monitoring by means of metadata, ecological regionalization and geostatistics using the UNESCO biosphere reserve Rhoen (Germany) as an example. Environmental Monitoring & Assessment, 114, 1/3, 461.