Research on Metadata Management System of Linkage Service of Scientific Data and Scientific Literature

Xiujuan Wang[1], Jiankui Chen[1], Xuerong Li[1,*]

[1]Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, Yantai, 264003, China

Abstract: In the data-intensive scientific research environment, the linkage of scientific data and scientific literature forms a complete body of scientific content. The literature and data serve scientific research together, which have become a hot issue of scientific research organizations. Starting from the metadata description elements of scientific data and scientific literature, this paper summarizes and analyses the association models of author association, keyword association and subject category association based on metadata description. On this basis, this paper describes the metadata management system architecture and system functions of linkage service of scientific data and scientific literature, providing some references for the relevant researchers.

1 Introduction

In the traditional academic exchange system, scientific research personnel regard scientific research literature as the most important scientific research achievement. However, literature itself cannot fully meet the needs of scientific research in-depth and innovative. Its supporting data also need to be related to the storage and sharing. As the scientific data attracts more and more attention in the whole scientific research process, scientific data sharing, opening and management has become a hot research topic in the fields of scientific research and information organization. Academic resources are no longer confined to traditional publications such as periodicals and monographs. The demand for scientific data and scientific research records has become the trend of academic information demand of scientific research organizations. Researchers believe that it is necessary to link academic documents with their supporting data. The association service between scientific data and periodical literature is an important way to change from traditional publications to convenient access to relevant scientific data. It has been studied and tried by many database providers and scientific data warehousing. The association between scientific data and periodical literature is of great significance to the acquisition and sharing of scientific data, the reuse and innovation of scientific data, the evaluation and evaluation of periodical literature and the transformation of academic exchange system. Scientific data refers to the original basic data produced by various scientific and technological activities and the data sets and related information processed according to different needs. As a kind of information resource, the specific formats and types of scientific data include: observation simulation data; classification glossary; mathematical expression; molecular, chemical, genetic expression; structure, physics, computational model; tables, charts, maps, pictures; field and experimental notes. It has become an urgent problem to improve the efficiency of scientific research activities by effectively associating scientific data with existing scientific research information support systems based on scientific and technological literature. If the scientific literature and scientific data can be interrelated to form the whole content of scientific research, the literature and data will be able to better play their value to support scientific discovery.

2. Metadata of scientific data and scientific literature

2.1 Metadata overview

Metadata is a kind of structured data that provides information resources or data, which is a structured description of information resources. Its function is to describe the characteristics and attributes of information resources or data itself. Metadata is descriptive information about data, which should reflect as much as possible the characteristics of the data set itself to facilitate the accurate and efficient use of data organization and utilization. Data content varies greatly in different fields. The fundamental purpose of metadata is to promote efficient use of data, and the other is to serve computer-aided software engineering. It mainly includes the description of data sets, the description of data items, owners, data sources, data production years, acquisition methods, etc. in data sets; the description of data processing information, such as processing tools, dimension conversion; the description of data quality, such as data accuracy, data integrity, spatial resolution.

*Corresponding Author: Xuerong Li

Rate, scale, data range, etc; Description of data conversion methods; Description of database foundation, update, etc. Through metadata, the database can be retrieved and accessed, which can effectively identify, evaluate and track the changes of resources in the process of use, realize the effective discovery, search, integrated organization and effective management of resources used, and further processing and secondary development of data.

2.2 Metadata of scientific data and scientific literature

The data content of different domains and different properties will be different. The content of metadata will also be very different when applied to research purposes. Scientific data metadata describes information as the main object, and its metadata features are as follows. The essence of metadata is the descriptive information about data. The digital expression of information is the process of abstracting, abstracting and transforming the geospatial information of the real world into the computer world. Metadata description provides a detailed description of the process to promote the effective organization and management of geographic information. Metadata changes with the change of the object described, and its semantic expression requires high flexibility and high scalability. Spatial objects have very complex attribute structure, which carries a huge amount of information. There are intersecting and infiltrating relations between objects and among objects. Each subject has its own unique interpretation perspective. Therefore, the descriptive metadata information itself will become very complex. Diversity refers to the huge system of metadata and the diversity of metadata types. Different disciplines and different research purposes of the same disciplines will lead to metadata description systems that require a specific structure. By considering the use of object-oriented ideas such as inheritance and reuse, we can solve the relationship and unification of diverse metadata structures. Although scientific data repository and scientific literature repository belong to heterogeneous databases, they have consistency in metadata description, and their metadata description reflects the common substance. Therefore, the association between scientific data and metadata-based descriptions of scientific literature can be established from these generalities.

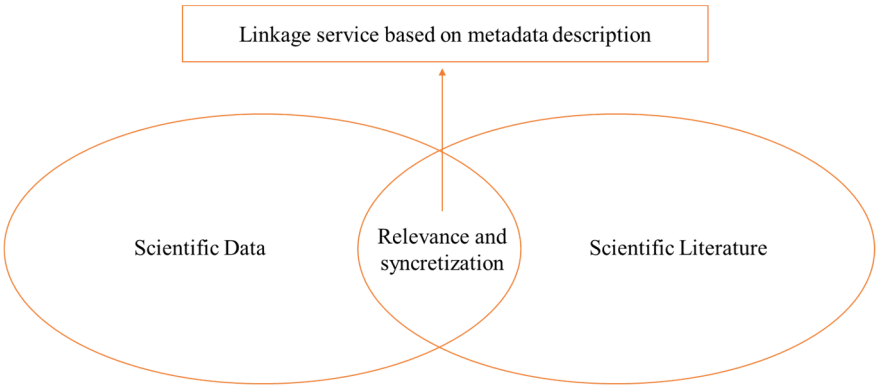

Figure 1. Linkage service based on metadata description of scientific data and scientific literature

3 Linkage analysis of scientific data and scientific literature

3.1 Linkage analysis of scientific data and scientific literature based on author description

With the development of experiment and observation methods and the importance of scientific data, researchers not only publish scientific papers in their research fields, but also submit some scientific data produced in scientific experiments and observations to corresponding scientific data centres. Therefore, many researchers not only create literature, but also produce many data resources in experimental, observational and other research. A certain author A publishes paper B and produces scientific data C. Because the researcher's research field is fixed and inherited, it can be concluded that paper B and scientific data C are related to a certain extent. Researchers' interdisciplinary and interdisciplinary research partnerships facilitate the emergence of new research results by linking seemingly unrelated data to the literature. A researcher A and other researchers B, C, D collaborate to publish papers or output data, and these collaborators B, C, D publish papers or output data in this field individually or/or in collaboration with other researchers E, F, G, thus expanding the network of researchers. There must be a link between research interests and results in this network, and there is also a link between the published literature and the output data. Therefore, cross-database retrieval technology can associate scientific data with scientific literature through author item metadata, and the scientific cooperation among these authors can help us associate several seemingly unrelated scientific literatures with scientific data, thus promoting the emergence of new scientific research innovation.

3.2 Linkage analysis of scientific data and scientific literature based on keyword

Keywords are the words that the author extracts from the title, abstract or text that best represent the subject matter of the document, that is, those words that are most important for revealing and describing the subject matter

of the document. We can get a glimpse of the general contents of the literature by keywords. Statistical analysis of literature shows that some internal links of literature keywords necessarily reflect some internal links between literature and its authors, which provide new materials for bibliometric research. For scientific data, the value of keywords lies in engineering data covering images, pictures, text, video and other forms. Therefore, the indexing of various forms of scientific data is a relatively complicated work, and through the analysis of the data content, the use of keywords can be any form of scientific data indexing. The development of keyword-based retrieval has a long history, and the technologies of keyword extraction and keyword indexing have been relatively mature. For users, the keywords are more self-contained. When searching, the users often choose the keywords which can accurately reflect their intentions as the search marks. There are two methods for indexing scientific data. Worker-worker indexing, i.e. data contributors or indexers, assign keywords to scientific data after analysing its contents. Automatically extract keywords from descriptive text of scientific data. Data description text is mostly provided by door data producers or data collators when they provide data. At present, most scientific data have data description. The descriptive text of scientific data is mostly composed of the introduction of relevant knowledge concepts, data acquisition methods, data selection and calculation methods. Automatic keyword extraction from descriptive text of scientific data can not only express the connotation of data more accurately, but also greatly improve the efficiency of indexing.

3.3 Linkage analysis of scientific data and scientific literature based on subject category

Subject classification is statistical, which provides a wide range of content and application prospects for us to conduct quantitative research on literature and reveal the correlation between documents. The classification number given in the scientific literature specifies the subject nature of the document from the content. If different documents have the same classification number, it indicates that they are studying the subject of the same subject property. Thus, the relevant scientific documents and their authors are linked from the perspective of subject classification, and the scientific literature group with the same subject property and the scientific literature group with the same subject property are formed. As the classification is based on the inclusiveness of disciplines by layers of cumulative expansion to reveal the relationship between disciplines. Therefore, the higher the rank of two or more scientific documents with the same classification number, or the finer the classification, the stronger the correlation between them. Among them, the more scientific documents of the same classification number, the greater the extent of their association. At the same time, the classification number can also closely link the author of the literature, forming a research direction with the same subject characteristics of the relevant author group. Therefore, the classification number of scientific literature and its authors are also of practical significance. Discipline refers to the knowledge system about the essential characteristics and laws of specific objects in the objective world. The general method is to use the classification number as the code of discipline classification. Each category concept in the taxonomy has its corresponding connotation, reflecting the essential attributes of things. Therefore, the information resources with the same category number have the same connotation and are related in content. If a scientific data and a document have the same classification number, then there must be a certain content correlation between them.

4. Metadata management system of linkage service of scientific data and scientific literature

4.1 System architecture

The metadata management system of linkage service of scientific data and scientific literature can be divided into four parts of standard layer, data layer, platform layer and application service layer.

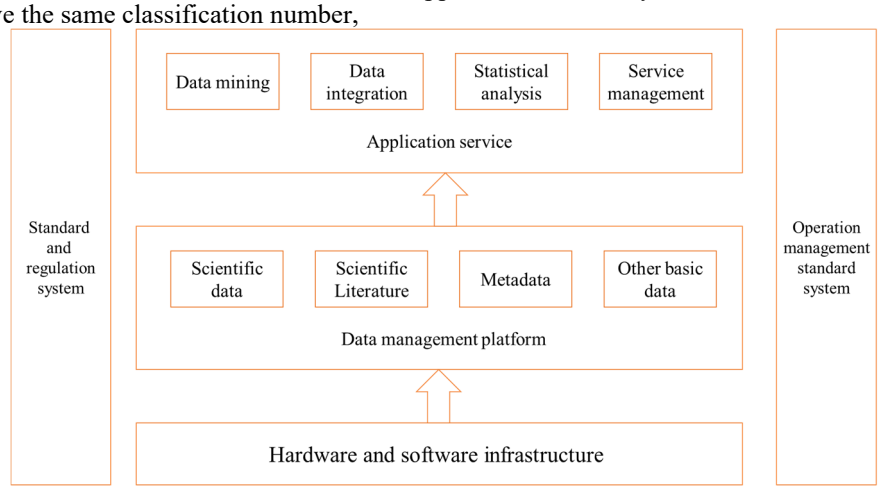

Figure 2. Architecture of metadata management system based on scientific data and scientific literature

(1) Standard layer: throughout the construction of the platform, including scientific data sharing and professional application service platform services related to a variety of standards and standardized text. In the

construction of the standard layer, the standard system suitable for scientific data sharing should be established and continuously improved based on the existing international and national standards.

(2) Data layer: The data base and source of shared services, including established scientific databases and metadata databases. The data layer consists of two main categories: entity data and metadata. Entity data includes all kinds of structured data as well as unstructured text, image and other data stored in this sharing platform. Metadata includes metadata of entity data and metadata information of user registration.

(3) Platform layer: software and hardware environment that ensures scientific data sharing. The data sharing platform is divided into two parts: foreground and background. The background is database management and application system. The foreground is network sharing platform system, sharing a unified database system and data services.

(4) Application service layer: A variety of applications based on scientific platforms, including data sharing services and analysis and processing services for users at regional, departmental, scientific and public levels.

4.2 System functions

System configurations include client and server configurations, information encryption, operation mode, database type and connection field configurations. User management can add, delete, audit, view related users, set passwords, change roles, permissions and other attributes, and freeze user accounts. The main purpose of roles is to assign privileges to the system, define the approval node in the workflow, improve the efficiency of multi-user privilege allocation, and reduce the workload of repeated privilege settings caused by account changes. This function module can flexibly set the role/user group information in the whole system. It can add, modify, delete and assign permissions to roles. It can flexibly set the user, role authorization scope, role data set permissions, set role permissions in batches, and role user group permissions in batches. Wait. According to different user levels and different internal management levels, data information can be effectively managed by setting different operation permissions in the organization. Guiding the system architecture through organizational management concepts prevents unauthorized people from going overboard and getting important information such as data and user information that they should not own. Menu management module is aimed at the developers and background administrators to manage the system for the purpose. The module is not for ordinary daily use. In the daily management and maintenance process, we do not recommend any modifications to this function module, which has ensured the normal operation of the system. Menu management is to organize and maintain the function menu of the system. We can set the validity of the function module, menu name, menu module movement, sorting position, operation permission association, user permission and role permission configuration.

5. Conclusions

The analysis of the relationship between scientific data and scientific literature based on metadata description elements is to help researchers acquire research data comprehensively and enrich their knowledge accumulation and integration. This paper constructs a metadata management system for the linkage service of scientific literature and scientific data. The main conclusions are shown as follows:

(1) The metadata of scientific data and scientific literature has the features of description, dynamics, complexity and diversity.

(2) The metadata of linkage service of scientific data and scientific literature include the association models of author association, keyword association and subject category.

(3) The metadata management system of linkage service of scientific data and scientific literature can be divided into four parts of standard layer, data layer, platform layer and application service layer. The system is a comprehensive data service platform which integrates metadata directory service, data service, function service and online data management.

Acknowledgements

The paper is the results of Engineering Project of Scientific Research Information Application of Chinese Academy of Sciences (Grant No. XXH13506-305) and Capacity Building Project of Literature and Information of Chinese Academy of Sciences (Grant No. ICP2017-5).

References

1. Nobre G C, Tavares E. Scientific literature analysis on big data and internet of things applications on circular economy: a bibliometric study[J]. Scientometrics, 2017, 111(1): 463-492.
2. Ziemann M, Eren Y, El-Osta A. Gene name errors are widespread in the scientific literature[J]. Genome biology, 2016, 17(1): 177.
3. Goodman A, Pepe A, Blocker A W, et al. Ten simple rules for the care and feeding of scientific data[J]. PLoS computational biology, 2014, 10(4): e1003542.
4. Howison J, Bullard J. Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature[J]. Journal of the Association for Information Science and Technology, 2016, 67(9): 2137-2155.
5. Beck F, Koch S, Weiskopf D. Visual analysis and dissemination of scientific literature collections with SurVis[J]. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1): 180-189.
6. Hirshkowitz M, Whiton K, Albert S M, et al. National Sleep Foundation's sleep time duration recommendations: methodology and results summary[J]. Sleep Health, 2015, 1(1): 40-43.
7. Altena A J, Moerland P D, Zwinderman A H, et al.

Understanding big data themes from scientific biomedical literature through topic modeling[J]. Journal of Big Data, 2016, 3(1): 23.

8. Grainger S, Mao F, Buytaert W. Environmental data visualisation for non-scientific contexts: Literature review and design framework[J]. Environmental Modelling & Software, 2016, 85: 299-318.
9. Dutra A, Ripoll-Feliu V M, Fillol A G, et al. The construction of knowledge from the scientific literature about the theme seaport performance evaluation[J]. International Journal of Productivity and Performance Management, 2015, 64(2): 243-269.
10. Swain M C, Cole J M. Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature[J]. Journal of chemical information and modeling, 2016, 56(10): 1894-1904.
11. Burley S K, Berman H M, Christie C, et al. RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education[J]. Protein Science, 2018, 27(1): 316-330.
12. Li J, Jovanovic A, Klimek P, et al. Bibliometric analysis of fracking scientific literature[J]. Scientometrics, 2015, 105(2): 1273-1284.
13. Tkaczyk D, Szostek P, Fedoryszak M, et al. CERMINE: automatic extraction of structured metadata from scientific literature[J]. International Journal on Document Analysis and Recognition (IJDAR), 2015, 18(4): 317-335.
14. Domingo J L. Safety assessment of GM plants: An updated review of the scientific literature[J]. Food and Chemical Toxicology, 2016, 95: 12-18.
15. Agapiou A, Lysandrou V. Remote sensing archaeology: Tracking and mapping evolution in European scientific literature from 1999 to 2015[J]. Journal of Archaeological Science: Reports, 2015, 4: 192-200.
16. Bornmann L, Haunschild R. Which people use which scientific papers? An evaluation of data from F1000 and Mendeley[J]. Journal of informetrics, 2015, 9(3): 477-487.