Research and Implementation of Scientific Data and Academic Literature Linking Service: Post-print

Authors: Huang Yongwen, Sun Tan, Zhao Ruixue, Guojian Xian, Li Jiao, Luo Tingting

Date: 2023-04-01T00:00:00+00:00

Abstract

[Purpose/Significance] To address researchers' growing needs for scientific data retrieval and discovery, this study enriches and improves the metadata of scientific data to achieve deep association discovery between scientific data and academic literature. [Method/Process] By analyzing and summarizing association service methods and practices both domestically and internationally, we propose a system architecture for scientific data retrieval and association services, and implement the collection and fusion of academic resource metadata, the enrichment and enhancement of scientific data metadata, as well as scientific data retrieval and association discovery services. [Result/Conclusion] Improvements in the quality of scientific data metadata can support deeper and finer-grained semantic association services between scientific data and academic literature, facilitating users' discovery of scientific data and associated academic literature.

Full Text

Research and Implementation of Linking Services Between Scientific Data and Academic Literature

Huang Yongwen[1], Sun Tan[2],[3], Zhao Ruixue[1],[3], Xian Guojian[1],[3], Li Jiao[1], Luo Tingting[1]

[1]Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081
[2]Chinese Academy of Agricultural Sciences, Beijing 100081
[3]Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing 100081

Abstract: *[Purpose/Significance]* To address researchers' growing demands for scientific data retrieval and discovery, this study enriches and improves scientific data metadata to achieve deep semantic linking between scientific data and academic literature. *[Method/Process]* By analyzing and summarizing domestic and international linking service approaches and practices, we propose an architecture for scientific data retrieval and linking services, and implement the collection and integration of academic resource metadata, enrichment and enhancement of scientific data metadata, and scientific data retrieval and linking discovery services. *[Result/Conclusion]* Improved quality of scientific data metadata can support deeper and more fine-grained semantic linking services between scientific data and academic literature, helping users discover scientific data and its associated academic literature.

Keywords: scientific data; academic literature; data retrieval; linking discovery
Classification Number: G253
DOI: 10.13266/j.issn.0252-3116.2021.23.013

With the rise and development of data-intensive research paradigms and data science, the supporting and enabling role of scientific data in scientific research, technological innovation, and evidence-based decision-making has become increasingly evident. At the national level, scientific data and academic literature resources have been classified by developed countries such as Europe and the United States as important components of national infrastructure. The European Union, the United States, Germany, and others have formulated relevant strategic plans and policies to promote the sharing and reuse of scientific data. Numerous publishers, funding agencies, research institutions, and academic societies have established scientific data sharing policies, with publishers explicitly requiring or recommending that authors submit supporting data along with their papers, assigning permanent unique identifiers (such as DOIs) to both literature and data. Data journals specifically publishing data descriptor papers have also emerged. Several important international organizations have launched action plans and standard frameworks, such as the European Open Science Cloud (EOSC) [1], Scholarly Link eXchange (Scholix) [2], FAIR Data Principles [3], DataCite Metadata Framework [4], and Elixir Interoperability Specifications [5], committed to building an open, reliable scientific data infrastructure and data sharing ecosystem that enables researchers to easily access and use scientific data, and calling for the creation of interconnection mechanisms between publishers and data repositories to facilitate access to and linking discovery of academic literature, scientific data, and other resources.

Meanwhile, researchers have gradually realized the importance of certain connections between scientific data and academic literature for improving research efficiency. According to Elsevier' s 2019 Trust in Research report [6], approximately 57% of researchers further examine the appendix data of literature. Linking academic literature and scientific data can enhance their discoverability and retrievability, and improve the transparency and reusability of research outputs. Currently, well-known international publishers, search engines, and data centers

have launched linking services for academic literature and scientific data, such as PubMed [7], Elsevier [8], Web of Science [9], Scopus [10], Dimensions [11], ELIXIR Data Platform [12], TAIR Arabidopsis Information Resource Platform [13], and ScienceDB, all of which provide services linking datasets to publications. Google Scholar [14], OpenAIRE [15], and RD-Switchboard [16] have built large-scale research graphs connecting academic entities/scientific communication entities such as literature, datasets, authors, institutions, and funders, thereby establishing a comprehensive and connected data ecosystem.

Although existing service systems have partially addressed the discoverability problem of scientific data, due to the decentralized and heterogeneous nature of scientific data and metadata quality issues, deep-level linking between scientific data and academic literature remains insufficient. Moreover, research in China on improving scientific data metadata quality and deep semantic linking between scientific data and academic literature is relatively limited, lacking exploration at the practical and applied levels. Therefore, this study focuses on the linking methods between scientific data and academic literature, analyzes foreign free linking services for scientific data and academic literature, and designs and implements a scientific data retrieval and linking service system to achieve semantic-level improvement of scientific data metadata and linking services with academic literature, providing a reference for Chinese library and information institutions to develop linking discovery services between scientific data and academic literature.

2 Related Research and Practice

2.1 Research on Linking Scientific Data and Academic Literature

In recent years, with the deepening concept of scientific data reuse and sharing, research on linking methods and relationship construction between scientific data and academic literature has gradually increased. Yang Ning et al. [17] categorized linking methods into active linking and passive linking, with active linking further divided into metadata-based linking, citation-based linking, and semantic-based linking. Jiang Enbo et al. [18] divided linking methods into hard linking based on form and soft linking based on content. Building on this foundation, we categorize research on linking methods into four main types: linking based on unique identifiers, linking based on citations, linking based on metadata, and linking based on semantic entities.

2.1.1 Linking Based on Unique Identifiers When scientific data and academic literature are stored and published, they are typically assigned digital object identifiers, enabling these research outputs to be individually discovered and cited. Unique identifiers for scientific data include DOI identifiers, data access numbers (e.g., database abbreviation: data identifier), ISLI identifiers, Handle identifiers, PURL identifiers, URN identifiers, ARK identifiers, CSTR identifiers, etc., among which DOI and data access numbers are the most commonly used unique identifiers for linking scientific data and academic literature.

Tu Yong et al. [19] and Sun Wenjia et al. [20] discussed the linking methods and key technologies for linking scientific data and academic literature based on DOI. Zhu Jiang et al. [21] studied the linking of academic literature and scientific data based on the ISLI standard. The German National Library of Science and Technology [22] has also actively explored linking between literature and scientific data based on DOI and linking literature and scientific data based on ORCID.

2.1.2 Linking Based on Citations Research activities generate large amounts of diverse scientific data each year, which are used by different groups and researchers and cited in publications, creating links between academic literature and their supporting data. This relationship not only makes scientific data reusable but also creates associations between academic literature and scientific data. Some scholars have studied citation-based linking methods, identifying and extracting datasets from literature in specific fields. Guo Xuewu [23] divided citation-based linking into three forms: direct citation linking, co-citation linking, and extended citation-based linking. Zhang Xin et al. [24] took the high-energy physics field as an example to study a literature-data linking algorithm based on citation probes, discovering implicit relationships through association degree calculations. Since authors often cite scientific data in the main text, some scholars have explored identifying scientific data from full-text papers. N. Riedel et al. [25] used text-mining algorithms to detect scientific data citations and availability statements in biomedical literature. L. L. Hou et al. [26] proposed a dataset entity recognition model MDER to extract cited and mentioned datasets from full-text content, validated in the computer science field. B. Ghaemi et al. [27] used a semi-automatic method to identify scientific datasets cited in social science literature.

2.1.3 Linking Based on Metadata Metadata-based linking between scientific data and academic literature primarily utilizes similarities in external and internal features of scientific data and academic literature to establish relationships. Sun Zhiru et al. [28] analyzed the similarities between data descriptions and literature descriptions in the bioinformatics field and proposed four linking methods: hard linking, neighbor relationship-based linking, data clustering-based linking, and topic-based linking. Huang Xiaojin believed that metadata association patterns between scientific data and academic literature include author linking, discipline classification number linking, and keyword linking [29], and extracted metadata items expressing content features from data and literature metadata descriptions to calculate relationships between data and literature using vector space models [30]. He Shuyue et al. [31] analyzed the similarities and differences in metadata between astronomical scientific data and academic literature and explored the feasibility of linking them based on data mining techniques.

2.1.4 Linking Based on Semantic Entities This approach mainly achieves linking between scientific data and academic literature from a semantic content perspective. Semantic entities refer to key concepts, terms, or entities (such as species names, gene names, protein names, chemicals, diseases, etc.) included in scientific data metadata descriptions. Methods for semantic entity recognition in scientific data mainly include author annotation and automatic text mining [32]. Sun Wei [33] used a faceted classification-based description method for fine-grained description of scientific data entities and verified semantic linking between scientific data and academic literature entities in the agricultural wheat breeding field. Ding Pei [34] proposed a fine-grained content semantic association model for academic literature and scientific data and verified the feasibility of entity recognition based on ontology for data-literature linking. T. Clark [35] proposed two complementary biomedical literature-data integration models based on entities and arguments. H. Cousijn et al. [36] established literature-data relationships based on table content, biological entities, and other methods. I. J. Aalbersberg et al. [37] used text mining and term extraction techniques to mine semantic entities from full texts and establish semantic entity-to-data repository linking.

2.2 Practice of Linking Services for Scientific Data and Academic Literature

Domestic practice in linking services between scientific data and academic literature is just beginning, so we mainly analyze foreign free linking service systems for scientific data and academic literature. DataCite [38] launched the DataCite Search data search tool in August 2015, providing one-stop data retrieval services and creating interconnection mechanisms between publishers and data repositories. Google [39] recognized the increasing importance of data and launched Google Dataset Search in 2018, elevating dataset discoverability to a new level. The Research Data Alliance (RDA) and World Data System (WDS) [40] collected links between data and literature from data centers, journal publishers, and research institutions, and launched the ScholeXplorer data-literature interconnection service. The U.S. National Institutes of Health (NIH) [7] relied on the "National Library of Medicine 2017-2027 Strategic Plan" action plan to develop and launch PMC-based data linking discovery services. OpenAIRE [41] established semantic relationships between datasets and publications across different disciplines through automatic reasoning, aggregating and linking datasets and publications to provide basic and additional services for the entire research process. Domain scientific data repositories such as Dryad [42], PANGAEA [43], and HEPData [44] have also achieved interconnection with literature. In addition to displaying basic scientific data information, they provide links to literature. The comparison of these free scientific data and academic literature linking service systems is shown in Table 1 .

As can be seen, mainstream search engines and data centers have begun to focus on collecting and aggregating scientific data, specifically retrieving and

discovering scientific data, and linking literature with scientific data. Data repositories have also actively developed linking services from scientific data to literature, particularly in natural science fields such as biology, physics, and medicine, where linking services between scientific data and academic literature are relatively mature, achieving deep semantic linking services between literature and data. In contrast, linking service practices in the social sciences are fewer. More publishers are cooperating with data centers and institutional groups to establish collaborative sharing mechanisms, actively achieving interoperability between data and literature, and emphasizing interoperability by providing standard data access interfaces (such as OAI-PMH API, Restful API, etc.) and following FAIR data principles to ensure that publicly archived data and literature are citable and linkable, effectively improving the retrievability, discoverability, interpretability, and reusability of scientific data.

3 Design of Scientific Data Retrieval and Linking Service System

In recent years, researchers' attention has expanded beyond traditional literature resources such as journal articles, conference papers, and technical reports to include scientific data as an important resource. Researchers often start from academic literature to discover clues about scientific data from content or references. Therefore, effectively linking scientific data and academic literature is crucial for improving research efficiency, enhancing scientific data reuse and sharing, and achieving deeper knowledge discovery. Drawing on foreign related service systems and focusing on the core problem of how to effectively retrieve scientific data and how to use the relationship between scientific data and academic literature to enhance discovery services, we mine semantic entities in scientific data and academic literature, enrich and improve scientific data metadata quality, enhance scientific data discoverability, and link scientific data with literature to achieve multi-level enhanced discovery and linking services that help users quickly discover scientific and technological resources. The architecture of the scientific data retrieval and linking service system is shown in Figure 1 [Figure 1: see original paper].

3.1 Academic Resource Metadata Collection and Integration

DataCite is the registration agency for scientific data unique identifiers (DOIs), and CrossRef is the registration agency for academic literature unique identifiers. Both have authoritative metadata collections for scientific data and academic literature. Therefore, we selected DataCite and CrossRef as data sources. We obtained metadata from DataCite and CrossRef through OAI-API, used Kettle tools for ETL data processing, and filtered data according to certain rules. We only parsed and converted fields for scientific data metadata containing linking relationships, constructing a scientific data metadata database and an academic literature metadata database containing over 3.76 million scientific data metadata records, all with DOI identifiers. Simultaneously, we parsed, ex-

tracted, deduplicated, and merged linking relationships between scientific data and academic literature and between scientific data and scientific data. Based on the data association model (see Figure 2 [Figure 2: see original paper]) and the relationship type controlled vocabulary defined by the DataCite metadata framework (such as "IsCitedBy," "IsSupplementTo," "HasPart," "IsDerivedFrom," etc.), we formed a relationship database containing over 8.58 million relationship records.

During data processing, we used a rule-based approach to identify and save scientific data citation records in CrossRef references. For example, when reference titles begin with "Data from:" or "Data for:" , we identified and extracted 6,483 data citations, adding the "IsCitedBy" relationship type to the relevant scientific data in the relationship database. Although the number of scientific data citations in references is currently relatively small, as scientific data citation becomes widely recognized by researchers and mandatory or recommended by publishers and funding agencies, more researchers will submit and share data, and the number of scientific data citations will gradually increase. By the end of 2020, over 13,000 journals supported data submission and sharing policies [45], promoting the establishment of scientific data and academic literature linking relationships in publication submission systems and upstream in the research lifecycle. Meanwhile, with the deepening cooperation between CrossRef and DataCite, the integrity and accuracy of citation relationships between scientific data and academic literature will be further ensured.

The relationship database formed in this study mainly includes relationships between scientific data and academic literature and between scientific data and scientific data. The top 10 relationship types by quantity are shown in Table 2 , where A and B in the relationship type descriptions refer to at least one being scientific data. As shown in Table 2, the most common relationship types are "IsPartOf" (28.9%) and "HasMetadata" (24.4%). "IsPartOf" indicates that scientific data A is part of scientific data B, while "HasMetadata" indicates that scientific data A has other metadata B. Next are "IsCitedBy" (8.0%) and "IsSupplementTo" (6.4%). "IsCitedBy" indicates that B includes A in its citations, while "IsSupplementTo" indicates that A is a supplement to B. "IsCitedBy," "IsSupplementTo," "References," and "Cites" are the main relationship types for linking scientific data and academic literature.

3.2 Scientific Data Metadata Enrichment and Enhancement

Describing and organizing scientific data is a prerequisite for sharing, retrieval, and utilization. Metadata can describe the content and format characteristics of scientific data. However, most current scientific data has minimal metadata, with incomplete descriptions or insufficient details, lacking subjects, classifications, etc. Based on an analysis of content feature metadata in scientific data, we used thesauri, ontologies, and other knowledge organization systems, along with scientific data titles, keywords, and abstracts (or descriptions), to achieve automatic classification, subject indexing, and semantic entity tag generation

for scientific data. We supplemented scientific data with thematic concepts, Chinese Library Classification (CLC) numbers, biological species entity tags, chemical substance entity tags, gene entity tags, and other information, enriching and enhancing the semantic metadata of scientific data to support faceted browsing and linking discovery services. The process of scientific data metadata enrichment and enhancement is shown in Figure 3 [Figure 3: see original paper].

3.2.1 Automatic Generation of Semantic Entity Tags Semantic entities mainly refer to meaningful entity names included in scientific data titles, keywords, abstracts, or descriptions, such as species names, chemical substances, gene names, and protein names. We primarily used dictionary-based methods for semantic entity extraction. First, we built domain entity dictionaries based on ISTI, Species2000, Uniprot, MeSH, etc. Then, we matched entities from these dictionaries against titles, keywords, descriptions, or abstracts. When matched, we identified and extracted the entity, marked its position, and automatically generated relevant entity tags for the scientific data. We mainly added chemical substance, biological species, and gene entity tags to scientific data. Examples of semantic entity tag extraction for scientific data are shown in Table 3 .

To establish semantic entity-based linking relationships between academic literature and scientific data, we also needed to identify and extract semantic entities from academic literature metadata in addition to scientific data metadata.

3.2.2 Automatic Classification and Indexing of Scientific Data In reference [46], we proposed a full-process automatic classification method based on a multi-factor algorithm combined with weighting strategies. This method has no domain or processing object limitations. Based on manual indexing experience and training corpora, we combined multiple influencing factors including classification number occurrence probability, keyword position weight, proportion of keywords under each classification number, and frequency of all keywords under the classification number to achieve automatic classification. By inheriting and reusing existing authoritative corpora (such as the English Super-Science Thesaurus STKOS and the Chinese Agricultural Science Thesaurus CAT) and building annotation corpora based on high-quality authoritative sources extracted from literature data—including terms, concepts, terminology, and other knowledge elements representing document content, plus discipline classification numbers revealing domain features—we established a subject term-classification number mapping database to ensure subsequent automatic classification accuracy. Without manual review and using only multi-factor algorithm calculations, the proposed automatic classification method achieved accuracy and F-values above 80% for random samples of multidisciplinary academic literature.

We used the multi-factor algorithm from reference [46] for automatic classification of scientific data. First, we segmented and extracted keywords from the

metadata information (title, abstract or description, and keywords) of the scientific data to be indexed to obtain thematic information. Then, we performed exact matching between the extracted keywords and keywords in the selected annotation corpora to obtain matched keywords and corresponding discipline classification numbers, calculating the frequency of keywords and their corresponding discipline classification numbers in the corpora. Finally, we performed weighted calculations based on discipline classification number occurrence probability, extracted keyword position weight, proportion of keywords under the matched discipline classification number among all keywords under that classification number, and frequency of all keywords corresponding to the discipline classification number under that classification number. We sorted discipline classification numbers by score and selected the top 5 as classification numbers for the scientific data. Examples of automatic classification of scientific data are shown in Table 4 .

3.3 Scientific Data Retrieval and Linking Discovery Services

Based on the semantically enhanced scientific data metadata database, we implemented scientific data retrieval, browsing, and linking discovery services (available at: http://www.agriknow.cn/nstl/datacite.html), supporting multi-angle browsing by publication date, type, biological species entity, chemical substance entity, gene entity, discipline classification (CLC), keywords, publisher, source (publishing institution), access license, funding agency, etc. We integrated this with academic literature retrieval, providing faceted limited retrieval such as "topic words" and "supporting data" to achieve bidirectional linking between scientific data and academic literature, supporting linking discovery and navigation services, and providing users with more research clues based on semantic entities.

We effectively integrated and linked scientific data-related literature and literature-supporting scientific data. On the search results page, users can browse academic literature associated with scientific data by selecting the "supporting data" facet on the left (see Figure 4 [Figure 4: see original paper]). When viewing academic literature details, users can directly link to datasets mentioned in the paper by clicking "related data" on the right side of the page (see bottom-right of Figure 5 [Figure 5: see original paper]), or when viewing scientific data details, they can click "related literature" to link to academic literature associated with that scientific data (see top-left of Figure 5 [Figure 5: see original paper]).

Scientific data can supplement and explain research results published as papers, helping users better understand the entire research process and enabling research reproduction and falsification. Finding and discovering data is a necessary prerequisite for reusing scientific data. We built a scientific data metadata database and relationship database based on DataCite and CrossRef data, designed and implemented a scientific data retrieval and linking service system with academic literature, achieved deep integration and linking of scientific and

technological resources based on external and semantic features, and used semantic entities to create relationships between literature and data, supporting deeper and more fine-grained semantic linking services that help users discover scientific data and associated academic literature through multiple pathways. However, this study has some limitations, such as using only thesaurus-based methods, which can cause ambiguity or incorrect classification for words with multiple meanings. For example, "Latex" refers to a chemical substance (rubber) but is also a data format. Therefore, future research will optimize semantic entity extraction methods using deep learning to avoid such problems. We will also conduct more in-depth research on linking services, such as identifying and linking data access control numbers in literature, recognizing more types of semantic entities (such as protein names, pest names, etc.), and calculating and reasoning relationships between scientific data and academic literature to discover deeper linking relationships.

References

[1] European Open Science Cloud [EB/OL]. [2020-10-20]. https://eosc-portal.eu/about/eosc.
[2] BURTON A, KOERS H, MANGHI P, et al. The Scholix framework for interoperability in data-literature information exchange [J/OL]. [2020-10-20]. http://www.dlib.org/dlib/january17/burton/01burton.html.
[3] WILKINSON M D, DUMONTIER M, AALBERSBERG I J J, et al. The FAIR guiding principles for scientific data management and stewardship [J]. Scientific data, 2016: 160018. https://doi.org/10.1038/sdata.2016.18.
[4] DataCite metadata schema [EB/OL]. [2020-10-20]. https://schema.datacite.org/.
[5] Elixir. Interoperability platform [EB/OL]. [2020-10-20]. https://elixir-europe.org/platforms/interoperability.
[6] Elsevier. Trust in research [EB/OL]. [2020-10-20]. https://www.elsevier.com/connect/trust-in-research.
[7] Discovering associated data in PMC [EB/OL]. [2020-10-20]. https://ncbiinsights.ncbi.nlm.nih.gov/2018/11/
associated-data-in-pmc/.
[8] Elsevier. Linking research data and research articles on ScienceDirect [EB/OL]. [2020-10-20]. https://www.elsevier.com/authors/tools-and-resources/research-data/data-base-linking.
[9] Web of Science. Data Citation Index [EB/OL]. [2020-10-20]. https://clarivate.com/webofsciencegroup/soluti
data-citation-index/.
[10] Scopus. Data linking [EB/OL]. [2021-03-10]. https://blog.scopus.com/topics/data-linking.
[11] Dimensions. Linked research data from idea to impact [EB/OL]. [2021-03-10]. https://www.dimensions.ai/.
[12] Elixir data platform [EB/OL]. [2021-03-10]. https://elixir-europe.org/platforms/data.
[13] GARCIA-HERNANDEZ M, BERARDINI T Z, CHENG H, et al. TAIR: a resource for integrated Arabidopsis data. Functional & integrative genomics, 2002, 2(6): 239-253.
[14] SULLIVAN D. An introduction to our knowledge graph and knowledge pan-

els [EB/OL]. [2020-10-20]. https://www.blog.google/products/search/about-knowledge-graph-and-knowledge-panels/.
[15] OpenAIRE. OpenAIRE-research graph [EB/OL]. [2020-10-20]. https://graph.openaire.eu.
[16] RD-Switchboard [EB/OL]. [2020-10-20]. https://www.rd-switchboard.org/.
[17] YANG Ning, WEN Yi, ZHANG Xin, et al. Research on linking high-energy physics scientific data and academic literature [J]. Library Science Research, 2019(1): 47-52.
[18] JIANG Enbo, PEI Yuxiang. Research on integration methods and examples of scientific literature and scientific data [J]. Knowledge Management Forum, 2019, 4(2): 69-79.
[19] TU Yong, PENG Jie. Research on integrating scientific data and academic literature based on DOI technology [J]. Digital Library Forum, 2007(10): 28-31.
[20] SUN Wenjia, CHANG E. Analysis of linking scientific data and academic literature [J]. Library Theory and Practice, 2017(3): 49-53.
[21] ZHU Jiang, LI Xinyi, JIANG Enbo, et al. Linking academic literature and scientific data based on ISLI standard [J]. Library Theory and Practice, 2020(5): 80-83, 91.
[22] KRAFT A, DREYER B, LOWE P, et al. 14 Years of PID services at the German National Library of Science and Technology (TIB): connected frameworks, research data and lessons learned from an analytical perspective [J]. Data science journal, 2017, 16(36): 1-10.
[23] GUO Xuewu. Research on linking scientific data and academic literature based on citations [J]. Information Science, 2014, 32(4): 59-62.
[24] ZHANG Xin, WEN Yi, YANG Ning, et al. Citation probe-based literature-data linking algorithm and application: taking high-energy physics as an example [J]. Information Studies: Theory & Application, 2019, 42(10): 151-156.
[25] RIEDEL N, KIP M, BOBROV E. ODDPub-a text-mining algorithm to detect data sharing in biomedical publications [J]. Data science journal, 2020, 19(42): 1-14.
[26] HOU L L, ZHANG J, WU O, et al. Method and dataset entity mining in scientific literature: a CNN + Bi-LSTM model with self-attention [EB/OL]. [2021-10-08]. https://arxiv.org/abs/2010.13583.
[27] GHAVIMI B, MAYR P, LANGE C, et al. A semi-automatic approach for detecting dataset references in social science texts [J]. Information services & use, 2016, 36(3/4): 171-187.
[28] SUN Zhiru, HAN Tao, YANG Wen. Analysis of relationships between bioinformatics scientific data and academic literature [J]. Library and Information Service, 2008, 52(2): 88-91.
[29] HUANG Xiaojin. Research on linking scientific data and academic literature based on metadata [J]. Information Studies: Theory & Application, 2013, 36(7): 27-30.
[30] HUANG Xiaojin. Research on linking scientific data and academic literature based on content features [J]. Journal of Modern Information, 2018, 38(1): 56-59.
[31] HE Shuyue, WEI Ren, WU Maochun, et al. Research on the application of

linking between academic literature and observational data in the astronomical field [EB/OL]. [2021-09-10]. https://d.wanfangdata.com.cn/conference/8469846.
[32] WEI Junchao. Research on linking practice between scientific literature and scientific data: taking Elsevier as an example [J]. Journal of the National Library of China, 2017, 26(3): 93-101.
[33] SUN Wei. Research and implementation of scientific data and academic literature linking discovery system [EB/OL]. [2021-09-10]. https://d.wanfangdata.com.cn/conference/7611510.
[34] DING Pei. Research on fine-grained semantic linking between scientific literature and scientific data [J]. Library Tribune, 2016, 36(7): 24-33.
[35] CLARK T. Argument graphs: literature-data integration for robust and reproducible science [EB/OL]. [2021-01-20]. http://www.isi.edu/ikcap/sciknow2015/papers/Clark.pdf.
[36] COUSIJN H, HAAK W, KOERS H. Finding better ways to connect research data with scientific literature [EB/OL]. [2021-01-20]. https://www.elsevier.com/connect/finding-better-ways-to-connect-research-data-with-scientific-literature.
[37] AALBERSBERG I J, KAHLER O. Supporting Science through the Interoperability of Data and Articles [EB/OL]. [2021-10-08]. http://www.dlib.org/dlib/january11/aalbersberg/01aalbe
[38] DataCite Search [EB/OL]. [2021-03-10]. https://search.datacite.org/.
[39] Google Dataset Search [EB/OL]. [2021-03-10]. https://datasetsearch.research.google.com/.
[40] ScholeXplore [EB/OL]. [2021-03-10]. https://scholeexplorer.openaire.eu/.
[41] OpenAIRE explore [EB/OL]. [2021-03-10]. https://explore.openaire.eu/.
[42] DRYAD [EB/OL]. [2021-03-10]. Our platform. https://datadryad.org/stash/our_{platform}.
[43] Elsevier and PANGAEA Link Contents for easier access to full earth system research [EB/OL]. [2021-03-10]. https://www.elsevier.com/about/press-releases/science-and-technology/elsevier-and-pangaea-link-contents-for-easier-access-to-full-earth-system-research.
[44] HEPData [EB/OL]. [2021-03-10]. https://www.hepdata.net/.
[45] STM. Research data share-link-cite [EB/OL]. [2020-10-20]. https://www.stm-researchdata.org/.
[46] LI Jiao, HUANG Yongwen, LUO Tingting, et al. Research on automatic classification based on multi-factor algorithm [J]. Data Analysis and Knowledge Discovery, 2020, 4(11): 43-51.

Author Contributions:

HUANG Yongwen: Writing and revising the paper
SUN Tan: Proposing the writing idea, finalizing the paper
ZHAO Ruixue: Revising the final version
XIAN Guojian: Data collection and processing
LI Jiao: Literature collection, investigation and analysis
LUO Tingting: Literature collection, investigation and analysis

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