FAIR Principle ：Make the Multimodal Data in Science & Technology Linkage Research Reliable

Chai Miaoling [1,2†], **Zhang** Xian [1,2*,†]

[1]*National Science Library (Chengdu), Chinese Academy of Sciences, China*

[2]*Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, China*

Abstract

The research uses the FAIR principle from the perspective of libraries to construct a reliable data selection and evaluation model for the study of the relationship between science and technology, in order to solve the problems of insufficient utilization of other data and unclear application logic, which mainly rely on paper and patent data in current research. Three issues are discussing in the paper: the format and interaction methods of multimodal data, the construction of indicator system, and the evaluation process. The result shows that the index construction method based on metadata interoperability is beneficial for the fusion of metadata layer, and use ontology to achieve semantic fusion, and identified scientific and technological relationships by multimodal data jointly.

Keywords

FAIR, multimodal data, metadata, reliable data, science and technology

1. Introduction and Related Works

The FAIR principles have become important in support of today's data-driven scientific research. The open access movement has expanded the methods for data acquisition and benefited researchers. However, due to the different intellectual property rights of data generated by scientific and technological activities, the lack of openness, incoherence, and sensitivity of the data, data acquisition is limited, and this affects the accuracy of science and technology correlation recognition. Therefore, this study introduces reliable data evaluation in science and technology linkage research to explore and evaluate the credibility of multimodal data participation in association recognition, striving to implement the principles into practice

[*] Corresponding author.

[†] These authors contributed equally.

chaiml@clas.ac.cn (Chai Miaoling); zhangx@clas.ac.cn (Zhang Xian);

0000-0002-5047-2167 (Chai Miaoling); 0000-0002-6297-1190(Zhang Xian).

https://doi.org/10.23106/dcmi.952401345

and unlock data’s potential in the research. [1-2]

Current works have explored the research under the FAIR (findable, accessible, interoperable, reusable) principle from the perspectives of scientific data management and cloud platform correlation, providing reference for the indicator. Wilkinson et al.(2016)[3] propose aims to enhance the machine's ability to automatically discover, utilize, and reuse data based on the FAIR principles, thereby strengthening its capacity for automated data search and utilization. Serena Bonaretti and Egon Willighagen (2019) [4] proposed to evaluate data fairness in life sciences, arguing that metadata standard frameworks and specific attributes of repository registries are conducive to achieving data fairness. Devaraju, Anusuriya et al. (2021)[5] discussed the FAIRsFAIR project's support for scientific data management under the FAIR principles, with its metrics primarily built upon the indicators developed by the RDA (Research Data Alliance) FAIR Data Maturity Model Working Group. Robert L. Grossman, et al. (2024)[6] introduced a complementary concept called SAFE, designed for cloud-based computing environments, which aims to support the interoperability of sensitive or controlled-access data (such as biomedical data) across two or more cloud platforms. Philippe Rocca-Serra et al. (2023) [7] developed a data management operations manual called the FAIR Cookbook specifically for the life sciences field, aiming to address the lack of practical guidance and help bridge capability gaps.

2. Research Methods

This study is a sub-research of the program, and the main target is to identify the linkage between science and technology using multimodal data. In the past, the linkage was usually identified via literatures and patents. With the popularization of open access more data could be acquired by librarian, but it is hard to find or accurately all the required resources. Under the circumstances, the project conducted research on available multimodal data types, and based on FAIR principle to create a Reliable Data indicator system by analytic hierarchy process (AHP). We define Reliable Data as trustworthy and reusable data that exists in the form of images, text, numerical values, etc. in the mining of science technology relationships, to help the users understand which kinds of data are missing and should be added from scientific research to the market, and also assist researchers in expanding the scope of data to enhance the credibility of research results.

To achieve this goal, the study focuses on the following three main parts: first, analyze the multimodal data modality and format; second, construct secondary indicators based on the FAIR principle and select metadata operations mapping it, with a perspective of machine automatic discovery and interoperability capabilities. Finally, score and evaluate based on 31 types of data.

https://doi.org/10.23106/dcmi.952401345

2.1 Multimodal Data Analysis

United Recognition of multimodal data requires understanding the data types, formats, and interaction methods firstly. Taking the agriculture field as an example, the multimodal data come from different stages of scientific and technological research, such as images, text, forms, electronic documents, etc. [8] These data exist in different formats, and have different interaction methods. Specific software or systems are used to interact with multiple datasets, please, check Table1.

Table 1

The Formats and Interaction Methods of Multimodal Data

Type	Format	Interaction Method
Image, pictures, etc.	tiff, jpg, png, GeoTIFF	Image Processing Technology
Text	txt, doc, docx，rtf	Microsoft, etc. related software
Form	xlsx, xls	Microsoft, etc. related software
oElectronic Documents	pdf	PDF reader
Standard Markup Language	xml, html, kml	Text editor or web design software
Binary files	bin	Related software
Packed File	tar, zip, tgz, gz	Compression software
Genome annotation information file	gtf, gff, gff3	Bioinformatics software or tools
related to geographic information	dwg, shp, FileGDB, gpkg, MapInfo File, bed, tsv	AutoCAD、ArcGIS、QGIS、MapInfo etc.
Datasets, Asset Bundles	rpb, accdb	Specific software or system, Microsoft Access etc.

2.2 Reliable Data Model Construction

The purpose is to select multimodal data from the metadata layer under the FAIR principle, ensure its interactivity, and evaluate it for subsequent research.

2.2.1Indicator System of Reliable Data

The research adopts the AHP to design a hierarchical indicator system, please check Table2. Firstly, a two-level indicator system is designed based on the FAIR principle and open data access. The first level indicator presents a progressive relationship, while the second level

https://doi.org/10.23106/dcmi.952401345

indicator is ranked from friendly to difficult. The importance of the data is ranked, with 1 being the most important. This study referred to the application of FAIR principles in existing scientific data management to construct secondary indicators.

1. Findable, it can be observed that some references were made to the open access principles for publications and research data in the IMI2 project guidelines for open access to publications and research data (2021).

2. Accessible Section, refer to FAIRness Maturity Indicators(MI)on discoverability and set the accessible secondary indicators corresponding to MI as metadata lifespan, access protocol, and access authorization.[9]

3. The interoperability section refers to the FAIR Cookbook's metadata interoperability in research data. The main consideration is the smooth interoperability of multimodal data, as well as the three ways of multi-source data interoperability: content mapping, ontology mapping, and identity mapping. [10] please, check Table 2.

4. Reusable, to consider the data useable via library's data classification (Primary, Secondary and Tertiary data).

Table 2

System of Reliable Data

Primary Indicator	Difficulty Level	Secondary Indicator	Difficulty Level	Description
Findable	4	White data	1	Have comprehensive and rich Metadata
		Grey data	2	Discoverable by both humans and machines but lacks complete and accurate metadata descriptions, can be transformed into white data by supplementing or enhancing the metadata, and allowing it to be combined with other datasets.
		Black data	3	Sensitive or Restricted Data
Accessible	3	Open access	1	Metadata Lifespan
		Internal access	2	Access Protocol

https://doi.org/10.23106/dcmi.952401345

		Inaccessible	3	Access Authorization
Interoperable	2	Direct use	1	Content mapping
		Need intermediary	2	Ontology mapping
		Difficult to interact with	3	Identifier mapping
Reusable	1	Primary data	1	Metadata records the original state of data
		Secondary data	2	Metadata helps users understand the current state of data and how it has transformed from its original state
		Tertiary data	3	Metadata facilitates ensuring the traceability and credibility of data

2.2.2 Model Construction and Evaluation

1. The Analytic Hierarchy Process (AHP) is used to determine the weights of the grading indicators. Establish an AHP model with a judgment matrix of A, where a_{ij} represents the relative importance of A_j, A_iFor example, if > 1, it indicates that the former is more important, and $a_{ij} = 1$ indicates that the importance of the two are equal., There is a matrix:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & \cdots & a_{2n} \\ \cdots & \cdots & a_{ij} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & \cdots & a_{nn} \end{bmatrix} \tag{1}$$

2. To establish the importance judgment of matrix elements. Develop the importance levels and their assigned values for factor comparison based on the 9-level criteria, and quantify them. 1 represents the comparison of two factors with equal importance; 3 represents a comparison between two factors, with the former being slightly more important than the latter, and so on. The third step is to calculate the weights of

https://doi.org/10.23106/dcmi.952401345

primary and secondary indicators, and multiply the weights of the two to obtain a comprehensive weight, which is then measured.

3. Based on the evaluation indicators, 31 commonly used scientific and technological data were scored to obtain core and non core datasets. The score range is set to 1-5 points. According to the ranking of scores, there are: Funds, Institutions, Market report, product data, Awards, Journal papers, Standard literatures, Invention patents, Academic monographs, Utility models, Experimental data, Production data, Conference papers, and Theses.

4. Empirical Research

From the perspective of librarian, the research acquired data via Open Access and verifies the model in the field of highland barley. The result show that in the core data sets, Rewards and Experimental data cannot be obtained by open environment yet. At last 12 types of data were obtained: funds, product data, journal papers, standards documents, invention patents, utility models, academic monographs, production data (seeds), conference papers, degree theses, institutions, and experts.

5. Discussion and Conclusions

Building a reliable data evaluation model based on the FAIR principle is beneficial for identifying the advantages and disadvantages of different data, and provides direction for stakeholders improving association recognition. Using the Interoperability of metadata as an indicator standard is conducive to achieving multimodal fusion, compensating for the insufficient data value mining. This model is advantage to the fusion in metadata layer, the reusable of literature and use ontology to fuse of multimodal data in semantic layer. Due to the disciplinary limitations, different results can be obtained if applied to different disciplines and research groups.

Acknowledgements

The work is supported by Chinese National Social Science Foundation Project. (Grant No. 18BTQ067). And we would like thank Professor Zhu Jiang in the test data download.

https://doi.org/10.23106/dcmi.952401345

References

[1] Kevin W. Boyack, Richard Klavans. Measuring science–technology interaction using rare inventor–author names[J]. Journal of Informetrics (2008) 2: 173–182.

[2] Dong Kun, Xu Haiyun, Luo Rui.et al. Review of the Research on Relationship between Science and Technology [J]. Journal of the China Society for Scientific and Technical Information, June 2018, 37(6): 642-652.

[3] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[4] BioRxiv[EB/OL].[2024-06-05]. https://www.biorxiv.org/content/biorxiv/early/2019/08/19/739334.full.pdf

[5] Devaraju, Anusuriya; Mokrane, Mustapha; Cepinskas, Linas.etc. From conceptualization to implementation: Fair assessment of research data objects. [J] Data Science Journal, Volume 20, Issue 1, Pages 1-14, 2021

[6] Robert L. Grossman., Rebecca R. Boyles, etc. A Framework for the Interoperability of Cloud Platforms: Towards FAIR Data in SAFE Environments. [EB/OL]. https://arxiv.org/pdf/2203.05097 January 22, 2024

[7] The FAIR Cookbook - the essential resource for and by FAIR doers. [J] Science Data. 19 May 2023

[8] Miaoling, C., Lin, H., Yunyue, R. A Review of Construction of Major Agricultural Open Scientific Data Resources[J]. Journal of Library and Information Science in Agriculture, 2020, 32(10), 25–34

[9] Bonaretti S , Willighagen E .Two real use cases of FAIR maturity indicators in the life sciences[J].Cold Spring Harbor Laboratory, 2019.DOI:10.1101/739334.

[10] FAIRCOOKBOOK[EB/OL].[2024-06-09]https://faircookbook.elixir-europe.org/content/recipes/interoperability/identifier-mapping.html

https://doi.org/10.23106/dcmi.952401345