Linked Data-Based Methods for Discovering Scholarly Relationships Among Entities in Institutional Repositories: A Postprint

Authors: Liang Meihong, Hu Xiaohui, Shu Pengfei, Liu Fei

Date: 2024-10-18T00:00:00+00:00

Abstract

This paper analyzes the types and characteristic attributes of entity academic relationships in institutional repositories, constructs a discovery method for such relationships based on linked data and elaborates its specific operational procedures, and conducts empirical validation using data from the Hohai University Institutional Repository. The results demonstrate that this method can achieve semantic aggregation and linking of institutional repository resources, provide linked data-based knowledge retrieval services, and precisely meet users' knowledge needs.

Full Text

Research on Methods for Discovering Academic Relationships Among Entities in Institutional Repositories Based on Linked Data

Liang Meihong, Hu Xiaohui, Shu Pengfei, Liu Fei

Abstract

This article analyzes the types and characteristic attributes of academic relationships among entities in institutional repositories, constructs a method for discovering these relationships based on linked data and explains its specific operational process, and conducts empirical tests using data from Hohai University' s institutional repository. The results show that this method can achieve semantic aggregation and association of institutional repository resources, provide knowledge retrieval services based on linked data, and accurately meet users' knowledge needs.

Keywords: Institutional repository; Linked data; Academic relationship; Knowledge discovery; Resource integration

Classification Number: G250.74

1 Introduction

Institutional repositories (IR) are important platforms for storing, managing, disseminating, and sharing academic knowledge resources of institutions, playing a significant role in promoting the sharing and exchange of research achievements, enhancing the influence of scientists and academic institutions, and improving the organizational management of academic output. Institutional repositories store various types and formats of academic achievement resources such as papers, patents, and research reports produced by an academic institution, which contain rich associative relationships. Discovering and revealing these relationships among resources helps achieve integrated organization and one-stop discovery of resources. Linked Data refers to connecting and integrating resources distributed across different times and spaces through linking technologies, and is an important technical means for institutional repositories to achieve semantic discovery of resources. Academic relationships are a type of associative relationship that exists in the scientific research process. Institutional repositories contain research entities that have academic relationships with each other, such as resources, scholars, institutions, and disciplines. Discovering and revealing these academic relationships among entities helps enhance the academic relevance of search results, improve user retrieval efficiency, and elevate the deep-level knowledge service capabilities of institutional repositories such as semantic retrieval, intelligent navigation, and scientific evaluation. This study explores methods for discovering academic relationships among entities in institutional repositories based on linked data technology, aiming to provide references for resource integration organization and knowledge discovery in institutional repositories.

2 Literature Review

2.1 Research on Academic Relationships Among Entities in Institutional Repositories

Currently, domestic and international research on entity relationships in institutional repositories has achieved certain results, mainly using existing conceptual models, computer technologies, and visualization tools. For example, establishing ontology models to achieve semantic association between entities [1-2], using association tables to store conceptual relationships among entities and conduct multi-dimensional associations [3], and using visualization tools to establish relationships between entities [4]. However, research results on the association and discovery of academic relationships among entities in institutional repositories

are relatively scarce. In the field of library and information science, academic relationship research mostly adopts methods such as social network analysis and multivariate statistics to quantitatively analyze the development trajectory of academic knowledge and academic exchange behavior [5-7]. The research team has conducted preliminary discussions on academic relationships among research entities in institutional repositories [8]179.

2.2 Research on Linked Data Applications in Institutional Repositories

Since its emergence, linked data has triggered extensive research and application. Its essence lies in tightly connecting various distributed and heterogeneous data through semantic associations to form an open, linked, and reusable data network. In 2007, the World Wide Web Consortium (W3C) launched the Linking Open Data (LOD) project and released the Linked Open Data cloud diagram. As of February 2023, the project has collected 1,594 linked open datasets [9]. In the library community, the Swedish National Library was the first to convert its national union catalog into linked data. Subsequently, the United States, the United Kingdom, France, and other countries have also released their library bibliographic data as linked data [10]. China has also achieved certain results in linked data research and application, such as the Shanghai Library releasing a series of knowledge bases based on linked data using its own resources [11].

Research on linked data applications in institutional repositories mainly focuses on three aspects: Application of linked data in institutional repository construction, such as Zhou Yu et al. [12] proposing a method for constructing institutional repositories based on linked data; Guo Weibing et al. [13] exploring the construction and service models of defense scientific research institutional repositories based on the basic principles of linked data. Application of linked data in data association and services of institutional repositories, such as Chen Jie et al. [14] exploring the association mechanism between scientific and technical report data and institutional repositories; Lin Jing et al. [15] studying extended services of institutional repositories based on linked data consumption. Application of linked data in resource discovery of institutional repositories, such as Wang Sili et al. [16] studying the application of linked data in semantic extension of institutional repositories; Du Pingping et al. [17] pointing out that linked data can be used to establish relational links for scientific research data in institutional repositories; Zhao Yiping [18] constructing a system for resource aggregation and knowledge discovery in institutional repositories based on linked data.

In summary, although academia has conducted research on entity relationships in institutional repositories and linked data applications, in-depth discussion on discovering academic relationships among entities in institutional repositories based on linked data is still insufficient. Therefore, this study explores methods for discovering academic relationships among entities in institutional repositories based on linked data technology, and uses Hohai University' s insti-

tutional repository as a data source to verify the feasibility and effectiveness of the method.

3 Method for Discovering Academic Relationships Among Entities in Institutional Repositories Based on Linked Data

The process of using linked data for academic relationship discovery is: Clarify the types of academic relationships among entities in institutional repositories and their characteristic attributes; Convert resource data into data conforming to the linked data format; Use appropriate association methods to discover academic relationships in the data and establish association links; Reveal the results of academic relationship associations through visualization methods to achieve semantic aggregation of institutional repository resources.

3.1 Types of Academic Relationships

In previous research, the research team used four types of entities—resources, scholars, institutions, and disciplines—as analysis objects, and found that academic relationships among different types of entities in institutional repositories include citation, affiliation, relevance, mentorship, collaboration, frontier hotspots, and interdisciplinary relationships [8]181. Specifically, resource entities have academic relationships such as citation, affiliation, and relevance—for example, a resource has a citation relationship with its references, and resources in the same discipline have relevance relationships. Scholar entities have academic relationships such as mentorship, collaboration, and citation—for example, the author of a dissertation has a mentorship relationship with their advisor. Institution entities have collaborative relationships, which include collaborations between first-level institutions, second-level institutions, and other institutions at different levels. Discipline entities have academic relationships such as frontier hotspots and interdisciplinary relationships. In addition, both resource entities and scholar entities have affiliation relationships with institution entities; resource entities, scholar entities, and institution entities all have research field affiliation relationships with discipline entities [8]181. The academic relationships among entities in institutional repositories are shown in Table 1 .

3.2 Linked Data Preparation Process

The prerequisite for achieving academic relationship association and discovery among entities in institutional repositories based on linked data is converting institutional repository data into data conforming to the linked data format. According to the concept of linked data and its four basic principles for publication (name resources with URIs; resource URI names can be found via HTTP URIs; when a resource URI name is found, useful information can be obtained; the obtained information contains links to more resources) [19], the linked data preparation process is divided into the following three steps:

First, data description and standardization processing. Resources stored in institutional repositories include manually uploaded data and machine-crawled data, which have different data structures and content. They need to be unified through standardized description, including constructing metadata sets for different types of resources, building authority files for scholars, institutions, and disciplines, and selecting appropriate schemes such as Dublin Core metadata (DC metadata) for data attribute description. It should be noted that institutional repositories usually already have some metadata sets based on specific metadata schemes, scholar dictionaries, institution dictionaries, etc., during construction. Therefore, when performing data standardization processing, it is necessary to consider reusing or mapping existing content to ensure data consistency and completeness.

Second, RDF conversion and URI naming. Linked data is data stored in Resource Description Framework (RDF) format and named with URIs [20]. Therefore, standardized institutional repository data should be converted into RDF format and named using URIs. RDF uses a triple format to express the relationships between resources, attributes, and attribute values, and carries semantic information, facilitating data identification and use [21]; URIs, as unique identifiers for internet resources [22], can reuse existing resources or customize identifiers according to relevant syntax rules. After conversion and naming, entities in institutional repositories obtain unique identifiers and can establish links. For example, the author attribute of a resource can be expressed as scholar:{, dc:creator, name}, and the institution affiliation attribute of a scholar can be expressed as institution:{, dc:institution, name}. Entities such as resources, scholars, institutions, and disciplines that have undergone linked data conversion in institutional repositories should have unique identifiers.

Third, linked data publication. There are currently multiple linked data publishing platforms. Institutional repositories should select appropriate platforms based on actual conditions such as resource scale and data update frequency, and publish standardized institutional repository data as linked data to provide resource guarantees for knowledge discovery services.

3.3 Academic Relationship Association Methods and Process

Based on the types of academic relationships that exist between entities in institutional repositories and the main characteristic attributes of corresponding resources, academic relationships among entities such as resources, scholars, institutions, and disciplines are discovered through sequential attribute value matching. When selecting characteristic attributes, priority is given to unique identifiers; in the absence of unique identifiers, relatively standardized attributes such as names, numbers, subject terms, and classification numbers are selected. The academic relationship association methods and process for institutional repository entities are shown in Figure 1 [Figure 1: see original paper].

At the resource level, academic relationships between resources are determined

based on attributes such as references, unique identifiers, affiliated projects, and disciplines. By matching whether the URI unique identifier attribute values of a resource and its references are the same, the existence of a citation relationship is confirmed. For example, in Figure 1, the identifier (unique identifier) attribute value of Resource 2 is the same as the reference attribute value of Resource 1, so Resource 1 and Resource 2 have a citation relationship. By matching attribute values such as resource source, type, and affiliated project, the existence of an affiliation relationship is confirmed based on whether the attribute values are the same. For example, in Figure 1, the project attribute values of Resource 1 and Resource 3 are the same, so Resource 1 and Resource 3 have an affiliation relationship. By matching attribute values such as author, institution, subject term, and discipline, the existence of a relevance relationship is confirmed based on whether the attribute values are the same. For example, in Figure 1, the subject attribute values of Resource 1 and Resource 4 are the same, so Resource 1 and Resource 4 have a relevance relationship.

At the scholar level, academic relationships between scholars are determined based on attributes such as author, advisor, and other authors. By matching the attribute values of authors and advisors of the same resource, the existence of a mentorship relationship is confirmed. For example, in Figure 1, Scholar 2 is the advisor (tutor) of Resource 4, and the creator (author) attribute value of Resource 4 is Scholar 1, so Scholar 1 and Scholar 2 have a mentorship relationship. By matching the attribute values of authors and other authors of the same resource, the existence of a collaboration relationship is confirmed. For example, in Figure 1, the creator (author) and contributor (other author) attribute values of Resource 5 are Scholar 1 and Scholar 3 respectively, so Scholar 1 and Scholar 3 have a collaboration relationship. By matching the attribute values of resource authors and reference authors, the existence of a citation relationship is confirmed. For example, in Figure 1, the creator (author) and contributor (other author) attribute values of Resource 5 are Scholar 1 and Scholar 3 respectively, and the creator (author) attribute value of Resource 6, which is a reference of Resource 5, is Scholar 4, so both Scholar 1 and Scholar 3 have a citation relationship with Scholar 4.

At the institution level, academic relationships between institutions are determined based on institution attributes. By matching institution attribute values of the same resource, first-level institution collaboration relationships are confirmed. Under the same first-level institution, second-level institution collaboration relationships are confirmed by matching second-level institution attribute values. For example, in Figure 1, Resource 6 has institution attribute values of Institution 1 and Institution 2, so Institution 1 and Institution 2 have a collaboration relationship. In Resource 6, Institution 2 has second-level institution attribute values of Second-level Institution 1 and Second-level Institution 2, so Second-level Institution 1 and Second-level Institution 2 have a collaboration relationship.

At the discipline level, academic relationships between disciplines are deter-

mined based on attributes such as discipline, keywords, and subject terms. By matching resource discipline attribute values, the existence of interdisciplinary relationships is confirmed. For example, in Figure 1, Resource 6 has subject attribute values of Discipline 1 and Discipline 2, so Discipline 1 and Discipline 2 have an interdisciplinary relationship. By matching resource keywords, subject terms, and other attributes, frontier hotspots in disciplines are obtained using methods such as social network analysis. At the level of different entity types, academic relationships between different entities are determined based on attributes such as author, institution, and discipline. For example, in Figure 1, the creator (author) attribute value of Resource 4 is Scholar 2, so Scholar 2 and Resource 4 have a contribution relationship; the institution attribute value of Scholar 4 is Institution 2, so Scholar 4 and Institution 2 have an affiliation relationship; Resource 6 has institution attribute value of Institution 2 and subject attribute value of Discipline 2, so Institution 2 and Discipline 2 have a research field affiliation relationship.

3.4 Application of Academic Relationship Discovery Results

Using academic relationship association methods can obtain academic relationships among four types of entities: resources, scholars, institutions, and disciplines in institutional repositories. By using data URI unique identifiers to point to each other, semantic linking of entities is achieved, and appropriate visualization tools are used to intuitively display linkable data with academic relationships, which is conducive to semantic aggregation and organization of resources and facilitates user browsing and information acquisition. Published linked data enables institutional repository resources to be associated with network resources, expanding the academic relationship network and providing strong support for academic exchange and diffusion.

4 Empirical Study of Academic Relationship Discovery Among Entities in Institutional Repositories Based on Linked Data

Hohai University's institutional repository comprehensively collects Chinese and foreign scientific research achievements of Hohai University' s faculty, and has built a metadata framework, author table, and institution table. This study uses this institutional repository as a data source to verify the scientificity, rationality, and effectiveness of the aforementioned construction method.

4.1 Data Acquisition and Processing

In Hohai University's institutional repository, 10 resource entity data items were selected as empirical research samples, and data processing was carried out according to the following process: Export data; Construct metadata sets and perform preliminary cleaning and processing; Based on the existing metadata framework, author table, and institution table in the institutional repository,

construct authority documents for scholars, institutions, and disciplines; Use the DC metadata scheme to standardize data description; Convert the standardized data into RDF format and name it with Uniform Resource Identifiers (URIs) to obtain data in linked data format.

4.1.1 Authority Document Establishment and DC Description First, construct metadata sets. Export the 10 sample data items into a structured two-dimensional table, where each column represents an attribute of a resource entity and each row represents all attribute values of a resource entity. Preliminary data cleaning is performed in Excel: split multiple author information in the other authors attribute into separate columns to ensure each attribute corresponds to only one attribute value; split first-level institution and second-level institution names in the institution attribute into separate columns and delete non-institution name content such as addresses and postal codes, retaining only second-level institution information of Hohai University; unify punctuation formats in Chinese and foreign language attributes; supplement missing data for important attributes, etc. Finally, a resource metadata set is constructed.

Second, construct authority documents. Based on the existing author table in Hohai University' s institutional repository and the faculty information table provided by the university' s personnel department, construct an authority document for scholars with Hohai University faculty and students as the main body, with missing content supplemented through research on the university' s official website and searches for author introductions in academic literature. Based on the existing institution table in the institutional repository, construct an authority document for institutions with Hohai University' s second-level institutions as the main body, with missing content supplemented through research on the university' s official website. Based on the existing discipline table in the institutional repository, construct an authority document for disciplines, using first-level disciplines published by the Ministry of Education as standardized names, and mapping Chinese Library Classification numbers, ESI disciplines, and WOS disciplines to the discipline names published by the Ministry of Education as aliases. The mapping between authority document attributes and fields in Hohai University' s institutional repository vocabulary is shown in Table 2 . Standardize and clean the data in the metadata set according to the constructed authority documents.

Third, standardize metadata description. The DC metadata scheme used in this study contains 15 core elements. According to Hohai University' s institutional repository' s existing metadata framework and academic relationship matching attribute requirements, metadata elements are extended to obtain a metadata description scheme (as shown in Table 3). Based on this scheme, DC metadata description is performed on empirical sample data and authority document data to unify and standardize data with different structures and formats.

4.1.2 RDF Conversion and URI Naming Convert the standardized data into RDF format according to the composition content of RDF triples to reveal the relationships between entities and attributes. Name the data according to URI naming principles for standardized description and data converted to RDF format. Data URI naming includes naming of four types of entities: resources, scholars, institutions, and disciplines, with naming results serving as attribute values for the entities' unique identifiers. Based on existing data URIs in Hohai University' s institutional repository, the naming mechanism is: prefix + unique identification number, where the URI naming prefixes for resources, scholars, institutions, and disciplines are respectively: "http://ir.hhu.edu.cn/Articles/Article_{Detail}.aspx?id=" , "http://ir.hhu.edu.cn/writer/rw_{zp}.aspx?id=", "http://ir.hhu.edu.cn/organ/jg_{zp}.aspx?id=" , "http://ir.hhu.edu.cn/class/db_{zp}.aspx?id=" . Finally, unified, standardized, cleaned, and described empirical sample data is obtained, providing data preparation for subsequent work.

4.2 Data Association Process and Association Results

Using the academic relationship association method proposed in this study, attribute value matching and reasoning are performed on empirical sample data to obtain academic relationships between entities and establish links.

Figure 2 [Figure 2: see original paper] uses the serial numbers in the standardized description data to represent entities, and shows the matching of academic relationships among entities in empirical sample data using examples with more types of academic relationships. At the resource level: Resource 1' s reference is Resource 8, and the two have a citation relationship; Resource 1 and Resource 2 have the same discipline attribute, both being atmospheric science and hydraulic engineering disciplines, and the two have a relevance relationship; Resource 1 and Resource 10 belong to the same National Natural Science Foundation project, and the two have an affiliation relationship. At the scholar level: In Resource 2, Scholar 2 is Scholar 1' s advisor, and the two have a mentorship relationship; Scholar 1 and Scholar 2 are both authors of Resource 1, and the two have a collaboration relationship; Scholar 1 and Scholar 12 are respectively authors of Resource 1 and its reference Resource 8, and the two have a citation relationship. At the institution level: Institution 1 and Institution 6 are both second-level institutions of Resource 1, and the two have a collaboration relationship. At the discipline level: Discipline 26 and Discipline 49 are both discipline attributes of Resource 1, and the two have an interdisciplinary relationship. Finally, academic relationship association results for all sample data are obtained.

Based on the academic relationship association results, semantic association and aggregation of entities with academic relationships are achieved through mutual linking and pointing of entity URIs, which improves the depth and breadth of academic relationship discovery and facilitates users' more convenient and efficient access to and utilization of resources in institutional repositories.

References

[1] Si Li, Chen Chen. Research on the Construction of a Research Data Ontology Based on BIBFRAME[J]. Journal of Information Resources Management, 2020(3): 110-117, 124.

[2] Nieto M A M, Diaz D A, Mora J, et al. An ontology-based approach to describe collaborative work by reusing and enriching data from an institutional repository[J]. Electronic Notes in Theoretical Computer Science, 2021, 354(S1): 129-139.

[3] Sun Yi, Pan Wei, Ma Lihua, et al. Design and Practice of a Multi-dimensional Association-based University Institutional Repository System Architecture[J]. Library Science Research, 2021(7): 35-43.

[4] Wu Zhiqiang, Zhu Zhongming, Liu Wei, et al. Research and Practice on Functional Extension of CSpace Knowledge Analysis and Visualization[J]. Data Analysis and Knowledge Discovery, 2019(3): 112-119.

[5] Ren Ruijuan, Pu Demin, Zhang Yuan. Knowledge Context Visualization Practice Based on Five-dimensional Academic Relationship Discovery[J]. Journal of Academic Libraries, 2016(1): 69-75.

[6] Feng Jingwen, Zhao Yong. Research on the Characteristics of Mentorship Relationships of Outstanding Scientists from the Perspective of Academic Genealogy—Taking Nobel Chemistry Laureate Lipscomb as an Example[J]. Information Engineering, 2020(6): 22-32.

[7] Yan Weiwei, Wen Xin. Attention and Collaboration: Analysis of Online Academic Relationship Networks and Behavioral Patterns of Scientific Research Users[J]. Information Theory and Practice, 2022(4): 75-82.

[8] Sun Qingyu, Liang Meihong, Hu Xiaohui. Research on the Academic Relationship Discovery System for Research Entities in Institutional Repositories[J]. Journal of Intelligence, 2022(11).

[9] The linked open data cloud[EB/OL]. [2023-02-20]. https://www.lod-cloud.net.

[10] Zhang Hailing. Research on the Linked Datafication of Library Bibliographic Data—Taking the German National Library as an Example[J]. Library Tribune, 2013(1): 120-125.

[11] Shanghai Library. Special Collections[EB/OL]. [2023-05-20]. https://www.library.sh.cn/resource?type=%E

[12] Zhou Yu, Ou Shiyan. Research on the Construction Method of University Institutional Repositories for Linked Data[J]. Library and Information Service, 2016(1): 105-113.

[13] Guo Weibing, Zang Lijuan. Construction and Service of Institutional Repositories Based on Linked Data Technology[J]. Journal of Ordnance Equipment Engineering, 2020(12): 275-280.

[14] Chen Jie, Han Fei, Wu Qian, et al. Research on the Data Association Mechanism of Scientific and Technical Reports[J]. Digital Library Forum, 2017(1): 46-50.

[15] Lin Jing, Chen He, Chen Juan, et al. Exploration of Extended Services in University Libraries Based on Linked Data Consumption—Taking Xiamen University Library as an Example[J]. Journal of Academic Libraries, 2020(3): 71-79.

[16] Wang Sili, Zhu Zhongming. Research on Using Linked Data to Achieve Semantic Extension of Institutional Repositories[J]. New Technology of Library and Information Service, 2011(11): 17-23.

[17] Du Pingping, Li Yuke, Meng Yong, et al. Research on the Association Organization of Research Data in the New Generation of Institutional Repositories[J]. Modern Information, 2018(12): 86-90.

[18] Zhao Yiping. Research on Resource Aggregation and Knowledge Discovery in Institutional Repositories Based on Linked Data[D]. Changchun: Jilin University, 2018.

[19] Linked data[EB/OL]. [2023-02-10]. https://www.w3.org/DesignIssues/LinkedData.html.

[20] Liang Meihong. Research on Association Methods for Bibliographic Linked Data[D]. Beijing: Institute of Scientific and Technical Information of China, 2016.

[21] W3school. RDF[EB/OL]. [2023-02-18]. http://www.w3school.com.cn/rdf/rdf_{intro}.asp.

[22] Uniform resource identifiers (URI): generic syntax[EB/OL]. [2023-02-18]. http://www.ietf.org/rfc/rfc2396.txt.

Author Information:

Liang Meihong (1991-), female, librarian, Hohai University Library, Jiangsu, Nanjing, 210098;

Hu Xiaohui (1983-), male, associate research librarian, Hohai University Library, Jiangsu, Nanjing, 210098;

Shu Pengfei (1993-), male, librarian, Hohai University Library, Jiangsu, Nanjing, 210098;

Liu Fei (1990-), male, assistant librarian, Nanjing Tech University Pujiang Institute Library, Jiangsu, Nanjing, 211200.

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