Semantic Annotation and Spatio-Temporal Search of Open Datasets

Xiaofeng Yan(✉), Jun Zhai, Yalin Zhou, and Jia Chen

School of Shipping Economics and Management, Dalian Maritime University, Liaoning, China
1182746197@qq.com, zhaijun@dlmu.edu.cn

Abstract. In today's big data era with the rapid development of the Internet, information and data are exploding, and this huge amount of data is available for us to process and use. There are 193 local government data open platforms in China, and the number of open datasets has reached more than 300,000. Reuse of these datasets is particularly critical to the development of social economy, politics, and production. Accurately finding the required datasets has become a new research hot spot. Improving the Web search discoverability of datasets has become one of the key initiatives to promote data flow and build a data ecosystem. Google's one-stop search engine fills the gap in dataset search, but there are two limitations: incomplete collection of datasets and the lack of spatial and temporal search. Therefore, this paper constructs a Spatio-temporal ontology based on the situation of the domestic open data platform and temporal and spatial properties. The dataset is semantically annotated based on the Spatio-temporal ontology. The paper crawl the annotated structured data and store it in CSV files, and then build a Neo4j dataset search system to conduct the cross-platform search of the dataset through temporal or spatial information. The research in this paper helps improving the discoverability, interconnectivity, and reusability of Spatio-temporal datasets, and promotes the formation of the data sharing ecosystem for China's open government data. It also helps the cross-border flow of open government data and its integration into the international data ecosystem, and has great reference value for the development of the search engine for open datasets in China.

Keywords: Open Government Data · Semantic Annotation · Dataset Search · Metadata · Ontology

1 Introduction

Since the launch of the U.S. data portal in 2009, the open data movement has emerged rapidly around the world. In 2013, the W3C launched the Data Action Plan [4]. By March 2022, DataPortals.org has collected 593 open data sites. Open data has become one of the main productive forces of modern society.

China also follows the world trend of open data. By October 2021, 193 local governments in China have established data open platforms. The development of open data in China has achieved remarkable results, and the number of government open datasets is also more than 300,000. Reuse of these datasets is crucial for the development of

K. Subramanian et al. (Eds.): CTMCD 2022, ACSR 99, pp. 737–747, 2023.
https://doi.org/10.2991/978-94-6463-046-6_84

social economy, politics and production. Any research needs data to back it up. After determining the research problem or model, the relevant personnel need to find the relevant datasets before the next work. In today's era of Spatio-temporal big data [3]. Rich Spatio-temporal datasets record human's activity information from different granularity, levels, and perspectives [2]. A Spatio-temporal dataset is a dataset with both temporal and spatial dimension properties. In order to improve the utilization rate of the dataset and facilitate the users of Spatio-temporal datasets to query the required dataset according to temporal or spatial information, it is necessary to add temporal and spatial elements to facilitate the accurate search of datasets. We can easily find the required information on traditional search engines, but we can't meet the need to find ideal datasets. Users can search for the datasets on each open data platform, but the access to the datasets is not convenient and comprehensive enough.

Google launched its new free dataset search engine in September 2018, helping researchers, journalists, and other users find needed data resources more easily [1]. Google's one-stop search engine fills the blank of dataset search, but there are two limitations: first, China's open data resources are not included; second, it cannot retrieve the dataset in the two dimensions of time and space, lacking the space and time search function.

2 Current Status of Research

Most of the foreign data open platforms use JSON-LD coding format to semantic annotate datasets, so that the dataset search engine can grab the dataset. China's government data open platforms have no semantic annotation of datasets. However some open scientific data platforms have made semantic annotation of datasets, such as the open research data platform and scientific database of Peking University.

In the metadata of datasets on foreign open data platforms, the construction of temporal and spatial elements is not perfect, and there is also the phenomenon of temporal and spatial metadata lacking values. China attaches less importance to the spatial and temporal attributes of datasets, and lacks the application examples of datasets. Compared with foreign countries, the temporal and spatial elements of the metadata on China's

country / Contrast item		external		internal	
		America	EU	Government Open Data	Open scientific data
open data platform		The US National Data Platform	The EU open data platform	Zhejiang data opening	Peking University Open research data Platform
URL		https://catalog.data.gov/dataset	http://data.europa.eu/	http://data.zjzwfw.gov.cn/	https://opendata.pku.edu.cn/
Semantic annotation		Microdata、JSON-LD	JSON-LD	not have	JSON-LD
The Dataset search engine		The Google Datasets Search Engine	not have	not have	not have
time	release time	√	√	√	√
	refresh time	√	not have	√	√
	Update frequency	not have	not have	√	√
	time frame	not have	not have	null value	null value
space	Release agency	√	√	√	√
	spacial scale	not have	√	null value	not have

Fig. 1. Comparison of semantic annotation and dataset search at home and abroad.

data open platform are more comprehensive. For example, the metadata of the dataset on the Data open platform of Zhejiang Province has release time, update time, update frequency, time range, release mechanism, and spatial scope, among which most of the values of time range and spatial range are mostly null. Datasets on the Open Research Data Platform of Peking University include release time, update time, update frequency, time range, and publishing organization. The element lacks spatial scope, and the value of time range is mostly null, as shown in Fig. 1.

3 Search Model for Datasets

Ontology is the basis of semantic annotation, and semantic annotation is the basis of dataset search. Under the guidance of the Spatio-temporal ontology, Authors add the temporal and spatial attributes to the metadata of domestic open datasets, then map metadata to Schema.org vocabulary and transform into structured data in JSON-LD encoding format. Next, the structured data are added to the source code of web pages. It completes the spatio-temporal semantic annotation. Finally, we built a database search system, crawl the annotated structured metadata information and store them as CSV files, which are imported into the database so that users can search the corresponding datasets according to temporal or spatial information across platforms.

3.1 Spatio-Temporal Ontology

Ontology is the basis of semantic annotation. This paper reuses the 11 existing ontologies (including dcat, dcat, dct, foaf, freq, gn, owl, rdf, rdfs, time, xsd and geosparql vocabulary).

In the Chinese context, since there is no unified standard for metadata in various government data open platforms, it is necessary to construct a Chinese standard terminology list of domain ontologies that can be commonly understood when studying and applying semantic annotation, in order to solve the current problems of inconsistent expressions of the same concept and unclear connotations of concepts (different meanings of the same name). The UML diagram of Chinese key terms list constructed in this paper is shown in Fig. 2.

Through the analysis and collation of dataset metadata and the Chinese standard glossary of terms, 17 core classes and their relationships, as well as 39 object properties and 37 data type properties are extracted. Among them, 7 classes are newly created in this paper's Spatio-temporal ontology. The namespace of the Space-temporal ontology in this paper is: http://www.semanticweb.org/86137/ontologies/2021/9/Spatio-Temporal.ontology, and the prefix is "st". The overall hierarchical diagram of the Spatio-temporal ontology is shown in Fig. 3.

3.2 Semantic Annotation

The key technical support for the dataset search engine to search the dataset is Schema.org. The semantic annotation of open dataset is to use the Spatio-temporal ontology to add type information to the data resources in China's open government

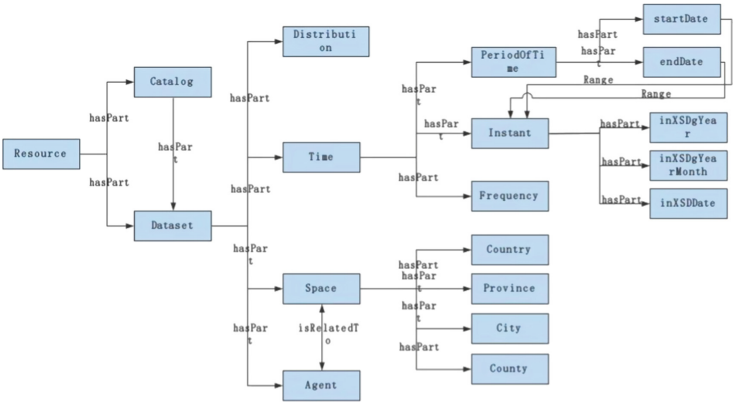

Fig. 2. Vocabulary of important terms of space-time ontology.

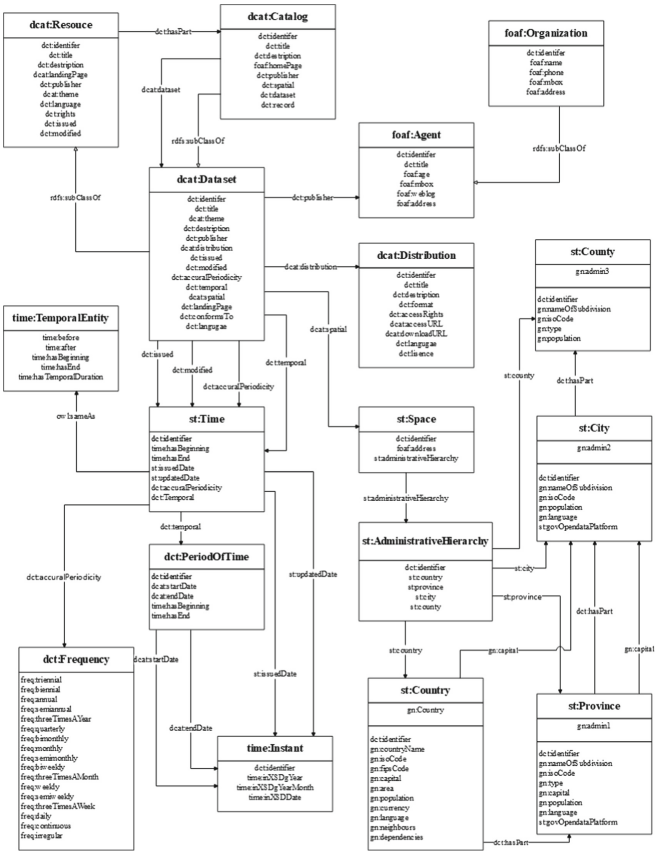

Fig. 3. UML diagram of the overall structure of spatio-temporal ontology.

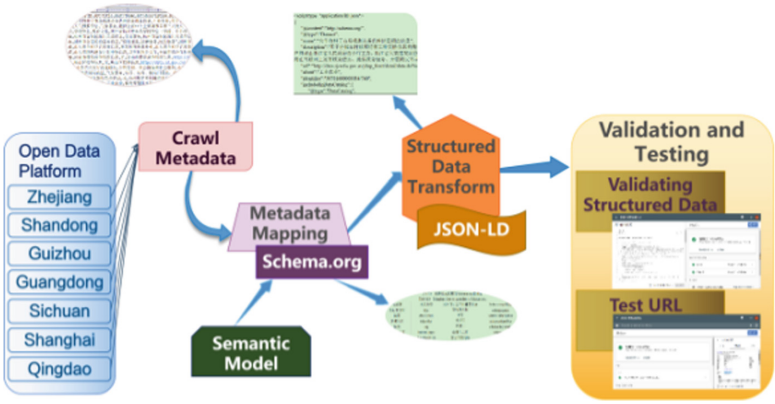

Fig. 4. The process of semantic annotation.

```
<meta name="SiteDomain" content="data.zj.gov.cn,data.zjzwfw.gov.cn">
<meta name="SiteIDCode" content="3300000068">
<meta name="ColumnType" content="数据开放">
<meta name="ColumnName" content="[数据详细页面]">
<meta name="ColumnKeywords" content="数据开放、数据集、数据应用">
<meta name="ColumnDescription" content="政府数据对外开放、提供利用。">
<meta name="ArticleTitle" content="农作物种子市场观测点备供种信息调查信息">
<meta name="PubDate" content="2021-12-01">
<meta name="ContentSource" content="省农业农村厅">
<title>数据详细页面</title>
<meta name="google-site-verification" content="7oThYJDOTu60SSJM8uNFISaQ6qnFk9J_2Xurk07_P_E" />
<script type="application/ld+json">
{
  "@context":"http://schema.org/",
  "@type":"Dataset",
  "name":"农作物种子市场观测点备供种信息调查信息",
  "description":"用于存储农村能源特有工种资格信息的数据信息，浙江省人民政府门户网站由浙江省人民政府办公厅主办，浙江省大数据发展管理
  "url":"http://data.zjzwfw.gov.cn/jdop_front/detail/data.do?iid=42085searchString=",
  "about":"工业农业",
  "identifier":"30701600003034/740",
  "includedInDataCatalog":{
     "@type":"DataCatalog",
     "title":"浙江-数据开放",
     "url":"http://data.zjzwfw.gov.cn/"
```

Fig. 5. Annotated page code.

data platform/website. We add structured data written in JSON-LD coding format into the original web source code, so that data resources can be identified, understood and included by the dataset search engine. The semantic annotation process is shown in Fig. 4, including "metadata crawling", "metadata mapping", "data structured transformation", "verification annotation code", "URL test".

The paper select the dataset "seed supply information survey information of crop seed market observation points" on the Zhejiang Provincial Government Data Open Platform (Zhejiang · Data Open) as an example to annotate. First of all, we crawl metadata of this dataset, add spatial and temporal properties, map the metadata of dataset to Schema.org vocabulary, then add the annotation code using JSON-LD encoding format to the web source code as shown in Fig. 5. The Glitch code tool can generate a new Web page URL (https://silly-locrian-seer.glitch.me), then perform the URL test, so far the semantic annotation of the dataset is completed.

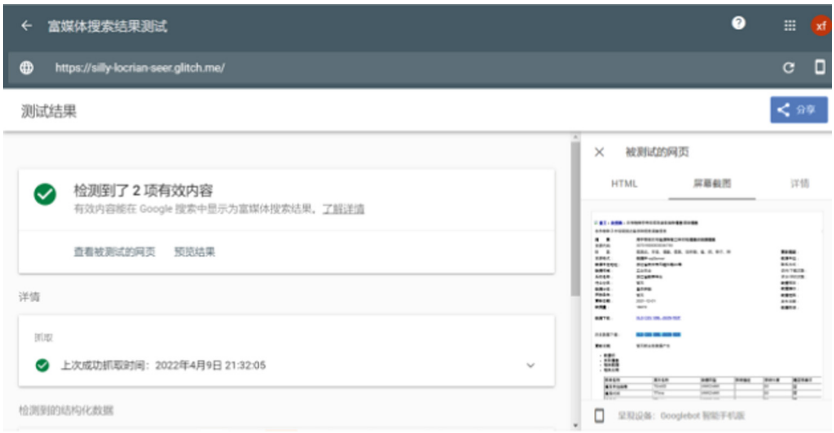

Fig. 6. Test generated web URL in rich media search results.

The URL of the web page generated by Glitch is tested in Rich Media Search Result, and the test result is shown in Fig. 6. 2 structured data items are detected, including a dataset and a publishing organization, indicating that the URL test passed.

3.3 Build a Diagram Database System

The data sources in this paper are selected from the five provincial government open data platforms ranked in the top 5 of the Open Forest Index in «Local Government Data Openness Report», and two municipal government open data platforms: Shenzhen and Guiyang. The metadata crawling, storage and refinement of 35 datasets on these 7 government data open platforms are carried out.

The process of dataset search: first through the annotated URL list, crawl the metadata records of all datasets, and store them as CSV local files. Then we improve the spatial and temporal attributes of datasets by combining the dataset metadata and the data itself. Finally, we import the data files in turn, and build the database system, as shown in Fig. 7.

According to the Spatio-temporal ontology, We crawl the data and store them into 15 entity class CSV files and 19 relationship class CSV files. Then we import these files into the Neo4j graph database. A total of 544 nodes and 617 relationships in the graph database are viewed as Fig. 8.

4 Spatio-Temporal Search of Datasets

The dataset Spatio-temporal search processes including “crawling structured data”, “perfecting spatiotemporal attributes”, “importing data files”, and “dataset spatiotemporal search” are shown in Fig. 9. The crawler traverses the list of URLs that have been annotated to obtain the metadata records and stores them as CSV local files. We Combine the dataset metadata and the data itself to improve the Spatio-temporal properties of the datasets, and add them to the CSV files. Then we build the dataset search system, import the data files in sequence, and perform the Spatio-temporal search of the datasets.

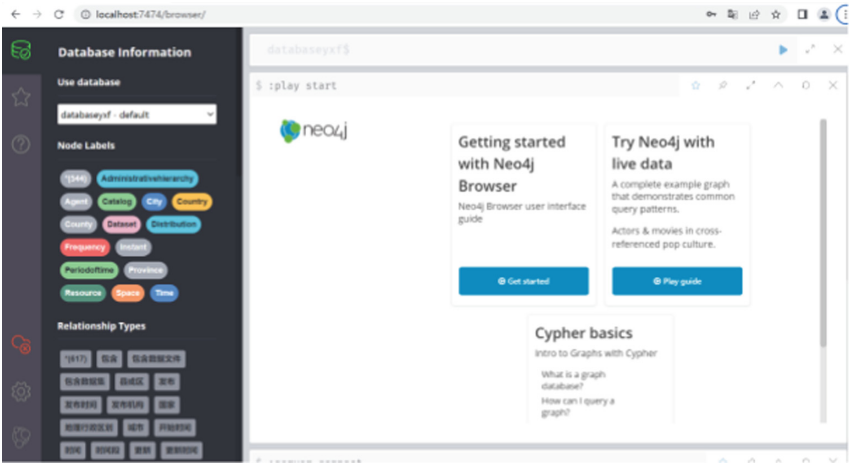

Fig. 7. Build graph database system.

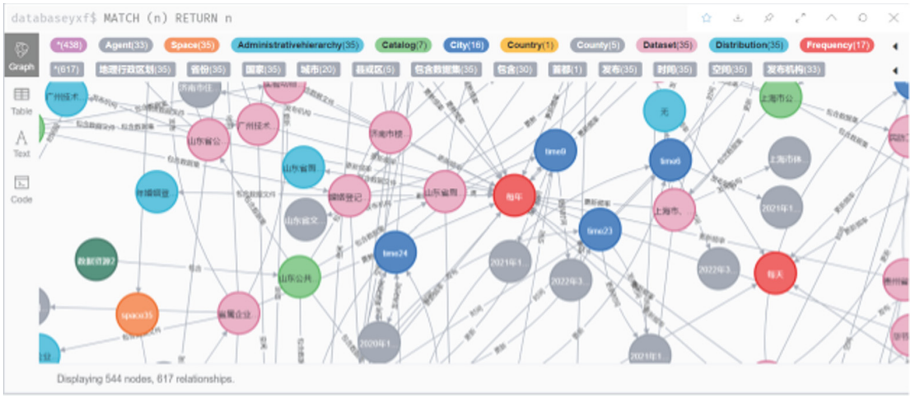

Fig. 8. View all nodes and relationships in the database.

4.1 Search Through Time Information

Users can query the dataset through the time information (including release time, update time, update frequency, and time range) of the dataset. For example, the user wants to query for datasets released in 2021. To match the dataset with "inXSDgYear" as "2021" by attribute, we enter the query script on the command line as follows.

```
MATCH (n:Dataset)-[:`Issued`]->
(m:Time)-[:`IssuedDate`]->
(x:Instant{inXSDgYear:'2021'})
RETURN n.title AS DatasetName,
n.landingPage AS DatasetURL,
x.tag AS IssuedDate;
```

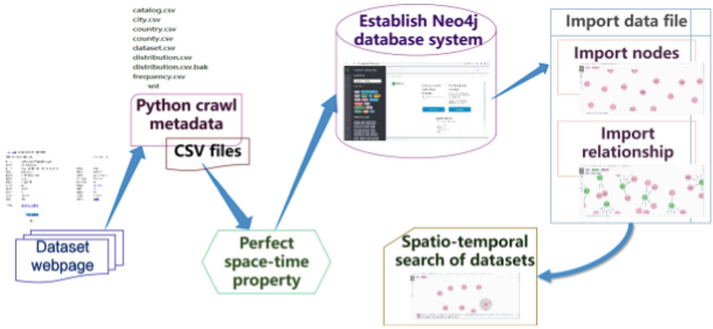

Fig. 9. The process of dataset search.

Fig. 10. Query datasets by issued date.

The query results are shown in Fig. 10, they show that there are 11 datasets issued in 2021, and the specific time of publication can be seen in the third column of the results. For example, the dataset of "Xiashan District Hospitality Information" is released on "November 23, 2021". In addition, you can jump to the homepage of the dataset through the link of the "dataset homepage" property.

4.2 Search Through Space Information

In this paper, we design the spatial hierarchy of this paper by referring to the hierarchy of geographic administrative divisions in China. The spatial hierarchy can be represented by the ASCII characters in Cypher query language as the following code.

```
(Dataset)-[:Space]->(Space)-[:Administrativehierarchy]->
(Administrativehierarchy)-[:Country]-(Country)-
[:hasPart]->(Province)-[:hasPart]->(City)-[:hasPart]-
>(County)
```

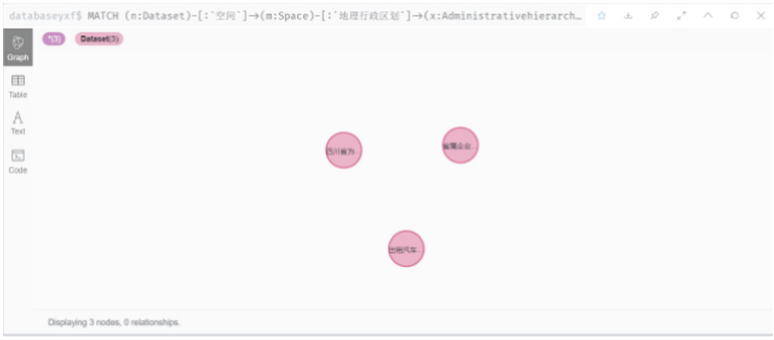

Fig. 11. Query datasets by city name.

Users can also query the dataset through the spatial information of the dataset (including the name of the country, province, city, county/district). For example, if a user wants to query a dataset related to "Chengdu", he needs to enter the following query script at the command line.

```
MATCH (n:Dataset)-[:`Space`]->
(m:Space)-
[:`Administrativehierarchy`]->
(x:Administrativehierarchy)-
[:`City`]->
(y:City{nameOfSubdivision:'Chengdu'})
RETURN n;
```

The query results are shown in Fig. 11, the results show that there are three datasets that belong to the city of Chengdu. You can click on the icon of the node and see the details of the node on the bottom edge of the result window.

4.3 Time and Space Search

The Spatio-temporal context of the dataset can also be maximum applied to query the dataset by combining temporal and spatial information. For example, if users want to query a dataset released in Shandong Province since 2018, it can be converted into a dataset with a release time greater than or equal to "2018", and the province of the dataset is "Shandong Province". Therefore, enter the following script at the command line:

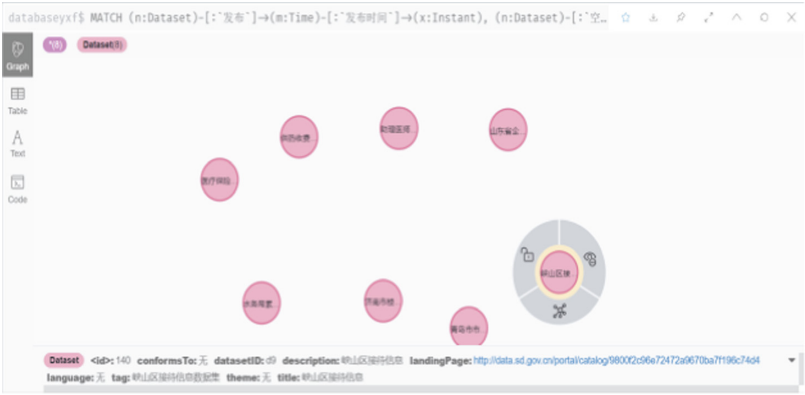

Fig. 12. Query datasets in combination with time and space.

```
MATCH (n:Dataset)-[:`Issued`]->
(m:Time)-[:`IssuedDate`]->
(x:Instant),(n:Dataset)-[:`Space`]->
(y:Space)-
[:`Administrativehierarchy`]->
(p:Administrativehierarchy)-
[:`Province`]->
(q:Province{nameOfSubdivision:'Shandong Province'})
WHERE toInteger(x.inXSDgYear) >=2018
RETURN n;
```

The query results are shown in Fig. 12, They show that there are 8 datasets released in Shandong Province since 2018, including those released in the provincial area of Shandong Province, as well as those in Jinan City, Qingdao City and Xiaoshan District of Weifang City. Each circular icon represents an instance dataset, and the details of the dataset can be expanded at the bottom of the query window. For example, the dataset "Xiashan Hospitality Information" contains the default id of "140" assigned by Neo4j graph database. The value of "datasetID" is "d9". The value of "description" is "gorge hospitality information". The value of "landingPage" is "https://data.sd.gov.cn/portal/catalog/9800f2c96e72472a9670ba7f196c74d4". The value of "tag" is "Gorge District Reception Information Dataset". The value of "theme" is "none". The value of "title" is "Information on hospitality in the Gorge District". In addition, You can jump to the homepage of the corresponding dataset through the link of "dataset homepage" property.

5 Conclusion

In the era of big data, the number of datasets has increased sharply, and it becomes urgent to accurately find the required datasets. In this paper, a Spatio-temporal ontology is created to semantically annotate the domestic open datasets. After the annotation is complete, we crawl the annotated metadata and store them in CSV files, and then import the data files into the Neo4j graph database, finally through the temporal and spatial information, we implements the cross-platform dataset search. The study in this paper helps to improve the discoverability, interconnection and reusability of datasets and fill the gap of Spatio-temporal search of datasets. At the same time, it Promotes the communication and interaction among open data platforms in China. It is convenient for researchers to find Spatio-temporal datasets quickly and precisely, and shortens the time for users to obtain satisfying datasets.

In the future, we will carry out semantic annotation of datasets all of our country, which can also be extended to datasets on the open science data platform. And then, users can search for all of the open datasets in China. In the next step, our team will explore the use of automatic annotation tools to improve the efficiency of semantic annotation. We will actively participate in the construction work of dataset search engine of our country. Next, we can also submit the annotated data resources to Google dataset search engine, which can promote the integration of China's open datasets into the international environment.

References

1. Brickley D, Burgess M, Noy N. Google Dataset Search: Building a search engine for datasets in an open Web ecosystem[C]. The World Wide Web Conference, ACM, 2019:1365-1375.
2. Di Yao, Chao Zhang, Jianhui Huang. Semantic understanding of spatio-temporal data: Techniques and Applications[J]. Journal of software, 2018, 29(07):2018-2045.
3. Eldawy A, Mokbel M F. The era of big spatial data[C]//2015 31st IEEE International Conference on Data Engineering Workshops. IEEE, 2015:42-49.
4. W3C. W3C Data Activity Building the Web of Data [EB/OL]. [2022-03-28]. http://www.w3.org/2013/data/.