Voovle: a Linked Data Search Engine for Scientific Data

Zhihong Shen, Yanfei Hou, Chengzan Li, Jianhui Li
Scientific Data Center
Computer Network Information Center, CAS
Beijing, China

Abstract—**Research activities are generating huge amounts of scientific data. This data deluge calls for search engines which can efficiently find out the required scientific data from a wide range of scientific data management systems for the users. The traditional document-oriented search engines are generally ineffective at discovering scientific data. This paper briefly introduces the background and requirement for developing scientific data-oriented search engines, and analyzes the issues and challenges are required to reflect upon and overcome for such search engines. It summarizes at the state of the art of the related work. And then in the main part of this paper, a linked data search engine for scientific data named Voovle is presented. It has been used in the Scientific Database project of Chinese Academy of Sciences. The workflow of Voovle is described emphatically. It consists of four processes including publishing, fetching scientific data and metadata, searching scientific data and discovering links among them. At last, the paper gives a glance at the application status of Voovle and the further work for improving it.**

Keywords- scientific data; linked data; search engine; data publishing; data fetching; data retrieval; VisualDB; Voovle

I. BACKGROUND AND REQUIREMENT

Research activities such as observations, experiments and investigation are generating massive amounts of scientific data. This data deluge is a big opportunity and, at the same time, is a big challenge for scientists. In order to exploit these huge volumes of data, some breakthroughs must be achieved in the fields of scientific data management [1]. One of the breakthroughs that must be achieved is to ensure efficient search and discovery across a wide range of scientific data types. The traditional web document-oriented search engines are not fit for them. Developing a scientific data-orientated search engine which can help people easily and effectively find the required data has become an important topic in e-Science.

For example, Scientific Database of Chinese Academy of Sciences (SDB) [2] is a database federation founded by Chinese Academy of Sciences (CAS) since 1983, in which there are multi-disciplinary scientific data accumulated through the course of research activities in CAS. By the end of the 11th Five-Year Plan of CAS, it had consisted of 51 databases and more than 500 data sets. A database is presented as a website in which users can search for data in it. By the end of 2010, the amount of online scientific data in SDB had reached about 141 Terabytes. The number of data files in it had been up to 26.76 million, and the number of data records stored in relational databases had been greater than 4.1 billion [3] .

During the 10th Five-Year Plan of CAS, We developed a search engine named *csdb-search* for the scientific data included in SDB. It is based on Nutch and oriented to web pages. At the same time, SDB provided a Google-based search portal for users. However, either *csdb-search* or Google could not get enough good retrieval efficiency. They searched the web pages but not the data in the databases. It often occurred that a lot of data related to the query could not be found out, because the web pages focus on how to present the data and how to make them human-understandable, but not on the content of the data or on how to make them machine-understandable. For example, the chemical compound structure is generally shown by a visual ActiveX control (or a Java Applet, an IFRAME, and other programs). It is obvious for the traditional web page-oriented search engines to be unable to search this kind of data. It leads to many related data excluded from the retrieval results. At the same time, a lot of irrelative retrieval results were found out because the adjunct content in the web pages (such as page titles, menus, footnotes, and etc.) undermines the retrieval precision.

II. ISSUES AND CHALLENGES

Contrast to document-oriented search engines, the following issues and challenges are required to reflect upon and to overcome when developing a scientific data-oriented search engine.

A. The diversity of data description formats and data organization methods

There are various storage formats of scientific data. In addition, there are different scientific data organization methods. Some scientific data are managed with RDBMS as records, and the others are organized in file systems as files. The data structure of the relational database is greatly influenced by the database designer's technical background and preference. The data files tend to be organized into file directories and the organization strategies also vary depending on the data curators' preferences.

The data description formats are very varied. Here it is illustrated by scientific data in the form of file. Their descriptive information (i.e., metadata) may be stored as the head part of the data file itself, or may be stored in a special description file, or may be stored as a database record. The descriptive elements used are usually different according to the different disciplines or subjects that scientific data are classified into. And the descriptive information is often bound with some special processing programs, and they could not be written and read by other software.

In summary, data description format and organization method is varied according to the discipline or subject, the

Jianhui Li is the corresponding author (E-mail: lijh@cnic.cn).

developer's or the data curator's technical background and preference. It is impossible to use a descriptive elements set such as Dublin Core Metadata Elements Set [5] to describe all kinds of scientific data.

B. Interfaces for scientific data open access

Just as mentioned above, SDB consists of 51 databases. Each database is built and maintained by one or more organizations. These databases (and the corresponding websites) are distributed in different geographical places. It is required to develop a standardized access mechanism which makes data accessed in machine-readable ways. This access mechanism should at least give solutions to the following problems.

a) The underlying communication protocol. For example, will standard HTTP services, web services, or other protocols be adopted?

b) How to identify a record and how to generate an unique identifier for each record?

c) How to package the request and the response when getting an appointed data record? How to represent each record? And how to package the attribute values? Whether to package them as binary content, or to package them through JSON or XML?

d) How to represent the relationships among data?

e) How to construct a search query and how to match it with the data?

C. Storage and retrieval of massive, heterogeneous scientific data/metadata

Differently from documents, database records usually include many attributes with different data types. The attribute value may be a string of characters, a figure, a date, a picture, a chemical compound structure, or even a binary stream. Therefore the search engine needs a storage system with strong capability which can store heterogeneous records without loss.

At the same time, the "massive and heterogeneous" character of scientific data/metadata means that it is more difficult to provide high-quality retrieval services for users.

D. Discovery and presentation of links among scientific data

There are links among scientific data. For example, the data about "red soil" in China Soil Scientific Database [6] is related to the data about the data about some subclass, the data about some places, and so on. It may be latently related to some data in another database such as Qinghai Lake Database [7]. The latently relative data may be the data about some birds (such as Anser indicus) in Qinghai Lake Database. Figure 1 shows the possible links among the scientific data mentioned above.

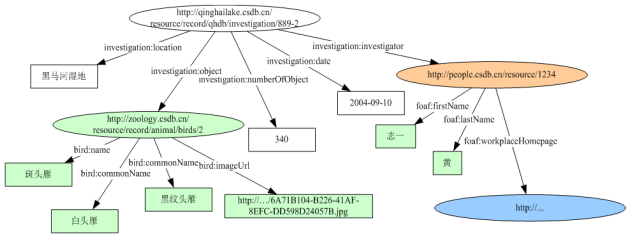

Figure 1. An example of the links among scientific data

So in the process of fetching and storing scientific data/metadata, it is required to try the best to avoid the loss of semantic linkage information and to display the complete data to the users. Because the links among resources distributed in different databases are latent, links discovery tools which can disclose this kind of links are required.

III. RELATED WORK

Scientific data storage formats are diversiform. Here are several examples of data storage formats: FITS format and PDS format used in the field of astronomy, DICOM format and PDB used in the field of biology, HDF format, NetCDF format and SDXF format used in multiple areas. In addition, there are many different metadata standards appropriate for scientific data in different disciplines, such as FGDC [8] and ISO 19115[9] used in the field of geosciences, Darwin Core [10] used in the field of biology, MCM[11] used in the field of medical science, EML used in the field of ecology, etc.

Among information description techniques, XML occupies an important place. Many data description formats are based on XML, such as MathML, CML, SMILES, etc. RDF and OWL make the ability to describe information improved. For example, RDF uses URI to identify each resource, and adopts the *subject - predicate - object* triple to represent each property. Such feature makes RDF become a description language used online naturally. Another advantage of RDF lies in links to other resources being expressed in RDF triples. It enables the data on the web more easily connected. Furthermore, in RDF(S) and OWL, data can be better organized with ontology.

Differently from the Interoperability protocols in the field of digital library such as Z39.50, OAI-PMH and OpenURL, many open access and acquisition interfaces of database are only open to a limited range of clients (generally to the systems located in the same LAN) . The SQL query interfaces are varied depending on the RDBMS products. Some RDBMS (for example, SQL Server) provide the function of creating web services directly. However, because web services are heavyweight, more and more work focuses on how to expose the database based on HTTP+JSON/XML. Some big IT companies such as Microsoft and Google also have developed their own open data protocols (such as OData and GData), and a lot of APIs and tools for data access and interoperability. Similar efforts includes SQL over HTTP (such as jtomyx, ChronicDB), restSQL, and so on. As to the exchange and sharing of data in the form of file, FTP and WebDAV are both standardized protocols. But a lower-cost, web-based and standardized interface is still required to be sought.

Tim Berners-Lee coined the term *Linked Data* in 2006 [12]. Linked Data has become the best choice among this kind of interfaces. Linked Data includes a set of techniques applied to the RDF data model that names all objects as URIs and makes them accessible via the HTTP protocol. Linked data emphasizes the linkages between data and the context information which are benefit for human being and machine's understanding [13]. Based on the concept of linked data, W3C initiated the Linking Open Data movement [14]. It has driven many data sets which are distributed in more than 200 domains published as Linked Data. By September 2010, LOD had covered about 25 billion RDF triples and about 395 million RDF links.

More and more data sets have been being published as Linked Data. Along with it search engines in linked open data context have been become a hot research topic. A lot of search engines of this kind have been developed, such as Swoogle, SWSE, Hakia, Squiggle, Falcons, Sindice, Watson.

Contrast to those researches above, our work focuses on applying Linked Data search engine techniques to scientific data. Some additional work are required according to the characteristics of scientific data, including developing VisualDB for publishing data, adopting the Sitemap protocol instead of SPARQL protocol, ranking the retrieval results based on the rank value, mining the links among data through links discovery tool, and so on.

IV. THE IMPLEMENTATION OF VOOVLE

The workflow of Voovle includes four processes: publishing scientific data/metadata, fetching scientific data/metadata, searching scientific data and discovering links among them. Figure 2 shows the modules involved in the processes, as well as relationships between them.

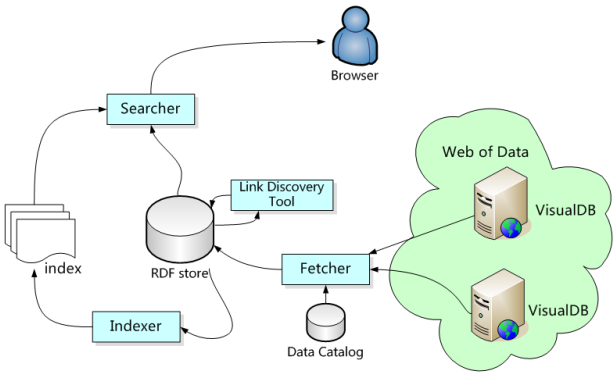

Figure 2. The Architecture of Voolve

A. Publishing Scientific Data/Metadata

The concept of scientific data varies in different contexts: it may mean a database, a data set, a data entity, a data item or a data value. Therefore we designed a hierarchical description model to distinguish different data in different granularity. The VisualDB[15] tool was then developed which adopts the model for organization, management and publishing relational records and data files. For relational records, VisualDB provides a model mapping interface to complete the mapping form E-R model to the hierarchical model with the help of user, such as mapping the column NAME of table PLANT into a property named dc:title. For data files, VisualDB enables automatically content metadata extraction and provides an interface for metadata annotation [16].

VisualDB enables publishing scientific data as Linked Data. Each relational record in the database is assigned an URI as follows:

<baseURL>/resource/record/<datasetName>/<tableName>/<itemIdValue>

In which *baseURI* is the service address of the database, *datasetName* is the name of the dataset, *tableName* is the table name which the record comes from, and *itemIdValue* is the primary key value of the record. As an example, a record which describes a "Bar-headed Goose" in Qinghai Lake database has the URI as follows:

http://www.qinghailake.csdb.cn/resource/record/qh/birds/2

Correspondingly, the URI of a data file is designed as follows:

<baseurl>/resource/file/<repositoryName>/<fileId>

In which *repositoryName* is the name of a repository in which the file is stored, *fileId* is a unique identity assigned by VisualDB. For example, a picture in the Plant Database has the URL as follows:

http://www.plant.csdb.cn/resource/file/repo1/1f62836e1c020c57011c020c57700000

When the browser requests the two kinds of URLs above, VisualDB will return the corresponding RDF graph of the data record or file metadata. A sourceUrl which refers to the location of the source file will be contained in such RDF graph for a file metadata.

B. Fetching Scientific Data/Metadata

By reading a directory of the database services address, the scientific data/metadata fetcher obtains the descriptions of schema including databases, data sets, data tables, data properties, and gets records and metadata of files in the same time.

For security and performance reasons, the fetcher does not connect the SPARQL backend of each data set using SPARQL query protocol. Instead, the Sitemap protocol [17] is adopted in the scope of the scientific database project. Sitemap index files are used for the big amount of records, the content of such an index file is shown as below:

```
<?xml version="1.0" encoding="UTF-8"?>
<sitemapindex xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <sitemap>
    <loc>http://www.qinghailake.csdb.com/birds-table.xml.gz</loc>
    <lastmod>2010-01-01</lastmod>
  </sitemap>
  <sitemap>
```

<loc>http://www.qinghailake.csdb.com/investigation-table.xml.gz</loc>

<lastmod>2010-01-01</lastmod>

</sitemap>

</sitemapindex>

Voovle fetches not only Schema Description Documents (SDD) but also Data Description Documents(DDD). A SDD includes the description information of a database (properties such as name, service unit, contacts, description, etc.), of data sets (properties such as name, collectors, description, acquisition time, acquisition methods, etc.), of data tables (properties such as name, description, etc.), and of data attributes (properties such as name, type, etc.). A DDD is often processed based on the original data record or data file metadata. It is therefore easy to associate each piece of data with the data tables and data sets in which it is stored, and the definition of all attributes in the record.

All of the original RDF graphs are saved in RDF / XML files as snapshots. All SDDs and DDDs are stored in a RDF store of Voovle, which is built on Jena TDB.

C. *Searching Scientific Data*

Voovle provides scientific data retrieval in three ways: by a SPARQL query, by a keyword fuzzy match and by a specialized query according to specific data sets.

Searching by a SPARQL query is suitable for advanced users, by accepting user input query, executing the query and displaying the results. One of the most common query expressions is:

SELECT distinct (?s) WHERE {?s ?p 'Anser indicu' }

In order to support keyword fuzzy match query, Voovle consider the acquisition time, the resolution and the linkage to other records of each DDD and give a rank value for each DDD, then perform the indexing and retrieval of data based on Lucene.

Since Voovle have harvested all SDDs for each database, it therefore provides professional retrieval functions for specific data set. For different types of properties, different matching operators are provided, such as "like" or "is" operators for text properties, and comparison operators like ">" or "=" operators for numeric and date attributes. Figure 3 shows a professional query interface for the seed oil measurement database in the China energy plant database.

Figure 3. The Query Interface for the Seed Oil Measurement Data set in China Energy Plant Database

D. *Displaying and Discovering Links among Scientific data*

Links exist among scientific data. VisualDB publishes foreign key relationships defined by E-R model as RDF links while publishing relational records. Voovle keeps all links lossless in the fetching process.

Figure 4 shows the interface for users when viewing the information of the "red soil" record, the links of its sub-class, geographical distributions and physical and chemical properties are listed below.

Figure 4. An Example of Links among Scientific Data

Since Voovle collects a large number of DDDs together to build a large RDF store, it is possible for it to discover new interlinks between different data sets. We developed and deployed a number of light-weighted link discovery tools according to specific requirements in Voovle. It has been proved that these tools find useful links and help users discover more knowledge. In the later, a configurable framework involving common similarity algorithms will be developed to help find more naive interlinks.

V. CONCLUSION

During the Eleventh Five-Year Plan of CAS, Voovle was deployed in both the SDB project and the "Basic Scientific Data Sharing Service System" project. Currently Voovle has a collection of 124 data sets included in 37 databases, the records amount to 5 million. It is proved that Voovle works well and provides stable searching service to users.

Further work include: to improve the sorting algorithm of the search results, to develop semantic mapping tools, to improve the link discovery framework, and to discover the links between the DDDs and scientific literature resources. Specialized retrieval systems will be also improved especially on the user interfaces to provide better service for researchers.

ACKNOWLEDGMENT

The work is supported by "Data Application Environment for Science Research" Project, CAS, 2005-2010(Project No: INFO-115-C01) and the "Specification and Service Platform, Basic Scientific Data Sharing Service System" Project (Project No: BSDN2009-17).

REFERENCES

[1] GRID 2020: A Vision of Global Research Data Infrastructure. Online at http://www.grdi2020.eu/Repository/FileScaricati/6bdc07fb-b21d-4b90-81d4-d909fdb96b87.pdf.

[2] Data Application Environment for Science Research Project. Online at http://www.csdb.cn.

[3] Data Resource Statistics system, Scientific Database of Chinese Academy of Sciences. Online at http://resstat.csdb.cn/.

[4] Cross-Domain Search System, National Science Library of Chinese Academy of Sciences. Online at http://crossdomain.las.ac.cn/SRW/frame/scisubject.jsp.

[5] Weibel S. The Dublin Core: a simple content description model for electronic resources. Bulletin of the American Society for Information Science and Technology, 1997, 24(1):9-11.

[6] China Soil Scientific Database. Online at http://www.soil.csdb.cn.

[7] Basic Database of Joint Research Center of Chinese Academy of Sciences and Qinghai Lake National Nature Reserve. Online at http://www.qinghailake.csdb.cn.

[8] Content Standard for Digital Geospatial Metadata — Federal Geographic Data Committee. Online at http://www.fgdc.gov/metadata/csdgm/

[9] ISO 19115:2003. Geographic information - Metadata.

[10] TDWG: DarwinCore. Online at http://www.tdwg.org/activities/darwincore/.

[11] Medical Core Metadata. Online at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243688/pdf/procamiasymp00002-0953.pdf.

[12] T. Berners-Lee, "Design issues: Linked data," Online at http://www. w3. org/DesignIssues/LinkedData. html, 2006.

[13] Linked Data FAQ. Online at http://structureddynamics.com/linked_data.html

[14] SweoIG/TaskForces/CommunityProjects/LinkingOpenData. Online at http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData.

[15] Zhihong Shen, Jianhui Li, Chengzan Li, Xing He, Xianming Su. VisualDB: Managing and Publishing Scientific Data on the Web. In Proceedings of CyberC'2011. pp.399~404

[16] Zhihong Shen, Yufang Hou, Jianhui Li. Publishing distributed files as Linked Data. In Proceedings of FSKD'2011. pp.1694~1698

[17] Sitemaps XML format. Online at http://www.sitemaps.org/protocol.html