DATA SCIENCE JOURNAL

Shen, Z et al 2015 Opencsdb: Research on The Application of Linked Data in Scientific Databases. *Data Science Journal,* 14: 4, pp.1-11, DOI: http://dx.doi.org/10.5334/dsj-2015-004

PROCEEDINGS PAPER

# Opencsdb: Research on The Application of Linked Data in Scientific Databases

Zhihong Shen[1], Jianhui Li[1] and Fang Han[1]

[1] Computer Network Information Center, Chinese Academy of Sciences (CNIC, CAS), Beijing, China
lijh@cnic.cn

This paper introduces OpenCSDB, a linked data application in the scientific database of the Chinese Academy of Sciences (CSDB) project. The background, applicability, implementation principles, software architecture, and application effects are discussed in detail. Linked data meets the needs of scientific databases as a low-cost, inclusive, adaptive, open access mechanism. Although there are new challenges, linked data can still be considered the best choice for scientific data. As proved by the achievement of OpenCSDB in the "Eleventh Five-Year" program, OpenCSDB will promote the sharing of scientific data and play a greater role in the "Twelfth Five-Year" program.

**Keywords:** Linked data; Scientific data; Database; Data records; Data files; Data publishing

## 1 Problems and Challenges

Researchers have accumulated a large amount of scientific data in long-term scientific research activity through scientific means such as observation, detection, testing, and surveying. The CSDB (Scientific Database Platform, Chinese Academy of Sciences 2011) project (http://www.csdb.cn), for example, involved 2 reference databases, 8 subject databases, 4 special topic databases, and 37 specialized databases during the "Eleventh Five-Year". These databases had accumulated more than 200 TB of scientific data, including 149.61 TB of online data, see Resource Statistics System, CSDB (2011) (http://resstat.csdb.cn/). This large-scale data deluge has brought great opportunities and challenges to scientific data management and data sharing (Gray, Liu, Nieto-Santisteban, et al., 2005). It has become the primary focus of the CSDB project in order to help scientific data researchers and e-Science applications to efficiently and conveniently find and consume data distributed in every corner of the web.

In the "Eleventh Five-Year", each database construction department in the CSDB project established a service website based on its individual database through a self-development approach or using a toolset of visual data management and publishing tools, VisualDB (Shen, Li, Li, et al., 2011). These websites followed a unified interface style, service mode, and user authentication standards and finally developed into a group of 51 scientific data service websites that provide a web-based data sharing service to end users. However, we also realize that current data sharing is based only on shared information on web pages. In fact, there has been increasing attention to the application of scientific data (database descriptions, data files, data records, etc.) themselves, rather than the presentation of HTML. Therefore, the CSDB project needs to find a mechanism for open access to scientific data (rather than web documents). According to the characteristics of scientific data as well as the CSDB project, open access mechanisms to scientific databases need to have the following characteristics:

(1) Inclusiveness of diverse scientific data. Scientific data from different disciplines differ in storage format and organization, and this diversity requires an open access mechanism to scientific databases to make them inclusive with different scientific data formats. In a scientific database, a large part of the data exists in a traditional file system in the form of data files while the other part of the scientific data is stored in different types of relational databases (RDBMS) in the form of scientific data records. The directory organizational strategy of data files and the structure design of data record tables are often dependent on the technical background and preference of data managers. On the other hand,

scientific data file formats, such as FITS, DICOM, PDB, PDS, HDF, NetCDF, SDXF format (Day, 2005), as well as scientific metadata formats, such as FGDC, ISO 19115, Darwin Core, MCM, see Scientific Data Formats (2012) (https://nf.apac.edu.au/facilities/software/IDL/docs-6.2/sdf.pdf) are complex and diverse; the open access mechanism to scientific databases needs to be inclusive enough to accommodate these description formats.

(2) Adaptability to diverse service environments. The CSDB project involves more than 50 database construction departments of the Chinese Academy of Sciences, each department varies widely in its data management systems, network environment, operating systems, and background knowledge of its IT technical support staff. As for relational data, for example, the data records are often stored, queried, and accessed by means of RDBMS. RDBMS data access protocols often have varied product types, such as Oracle, MySQL, SQL Server, and the versions of the database product vary greatly. Therefore, the open access mechanism to scientific databases requires a good security in different database products and other service environments.

(3) Semantics-supported and linked data. Scientific data have a natural semantic content, each table, such as birds, each property, such as title, in the scientific database, and the relationship between different concepts of the same meaning, such as the affiliation between "bar headed" and "bird", the equivalence relationship between "CO2" and "carbon dioxide". Taking the Basic Database of Joint Research Center of Chinese Academy of Sciences and Qinghai Lake Nature Reserve (http://www.qinghailake.csdb.cn) as another example, in addition to the inspection sites and the number of inspection objects, the investigated objects, such as bar-headed goose, need to be able to be linked to a description record in an animal database, including name, alias, bird photos, and the personnel involved in the expedition need to be linked to a record in the staff database. The open access mechanism to the science database must have a strong descriptive ability to express this correlation to show users the complete data.

(4) Promotion on the implementation level. Unlike literature resources, scientific databases currently lack comprehensive data publishing and data rights protection mechanisms. Scientific database resources are often in the hands of each individual institute research team. Ignoring the interests of the data in the data sharing process will bring greater resistance to data sharing; thus an open access mechanism for scientific databases requires non-mandatory, non-centralized features. In addition, this mechanism must be adequately lightweight and standardized so that it will not only make data sharing in a web environment more convenient but also will not cause high renovation costs or affect the existing web architecture and scientific database structure.

In summary, due to characteristics such as the diversity of formats and organization of scientific data, the differences in service environments and levels, the differences in semantics and relevance of data, the complexity of protecting rights and interests services, and so on, the CSDB project requires an inclusive, universal, associated and semantics supported, decentralized, and low-cost open access mechanism.

## 2 Related Research and Linked Data

In resource description technology, XML plays a significant role. Many scientific data description formats, such as MathML, CML, SMILES, and EML, use XML. The appearance of RDF (Resource Description Framework) and OWL (Web Ontology Language) enhanced the ability of resource description. For example, RDF uses a URI to identify each resource and a "subject - predicate - object" triple to represent each attribute. This design makes RDF a natural description language on the web. In addition, by using RDF (S) and OWL, people in the scientific data field can construct ontologies to organize their data.

In distributed data access technology, interoperability protocols such as Z39.50, OAI-PMH, and OpenURL dominate the field of digital libraries. Unlike in the library field, many scientific databases tend to open the SQL query interface of a database within a limited range (usually in LAN) based on a relational database management system, and their underlying protocols vary in different products, such as SQL Server and MySQL. Some databases (such as SQL Server) provide database web services. However, due to the complexity and heavy weight of web services, more and more studies focus on how to expose the contents of the database based on HTTP + JSON / XML. For example, Microsoft, Google, and other major IT companies have put forward their own open data protocols, such as Open Data Protocol (OData) (http://www.odata.org/) and GData (http://www.odata.org/), and provide a range of API and tools to support the access to data and their interoperability. Similar results also include SQL over HTTP (such as jtomyx, ChronicDB, etc.) and restSQL. As for the exchange and data sharing of scientific data files, FTP and WebDAV still dominate the important positions as standardized protocols in data sharing. However, we can see that these interfaces apply only to open access to file contents, and they cannot achieve perfect integration with other contextual information.

In this background, linked data has a very important significance. Tim Berners-Lee (2006), "Father of the Web", put forward the concept of linked data (http://www.w3.org/DesignIssues/LinkedData.html) in July 2006. As for its technical framework Berners-Lee (2006) (http://structureddynamics.com/linked_data.html), linked data is a collection of best practices. It uses the RDF data model, using URI (Uniform Resource Identifier) to name data entities to publish and deploy instance data and class data that can then reveal and obtain data through the HTTP protocol while emphasizing the interrelationships and communication of data and contextual information that benefit human-computer comprehension.

W3C's Linking Open Data Project (LOD) was officially launched in May 2007. Since its first three years, more and more data owners have been publishing their data in the form of linked data on the web. Through September 2011, LOD included 295 data sets, containing 31 billion RDF triples and 500 million RDF links, see SweoIG/TaskForces/CommunityProjects/LinkingOpenData (2010) (http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData).

In the field of library science, the Swedish Union Catalogue (LIBRIS) was the world's first union catalogue to release bibliographic data as linked data. Following LIBRIS, at least five international or national bibliographic data libraries standardized their data in linked data services. In May 2010, W3C established the W3C Library Linked Data Incubator Group (http://www.w3.org/2005/Incubator/lld/) to help increase global interoperability of library data on the web, by "bringing together people involved in semantic web activities—focusing on linked data—in the library community and beyond, building on existing initiatives, and identifying collaboration tracks for the future". In addition, well-known institutions such as the British Broadcasting Corporation (BBC), the New York Times, Reuters, and Best Buy, have adopted linked data in their everyday operations.

In the field of scientific data, linked data has been used as a standardized mechanism for open data in all disciplines. For example, Linked Life Data (Momtchev, Peychev, Primov, et al., 2009) integrates over 20 data sources, such as UniPort, PubMed, and Entrez Gene, to provide an integrated search and browsing service. The NCBO Resource Index (Jonquet, LePendu, Falconer, et al., 2011) applications provide a browsing service in biomedical resources to users through the utilization of more than 200 existing ontology of knowledge. The Diseasome Map (http://diseasome.eu/map.html) application integrates different bioscience data sources, generating diseases gene networks by associating known diseases with genetic disorders.

At present in China, research institutes studying linked data are concentrating on the field of digital libraries, such as the Shanghai Library Digital Library (Liu, 2010a; Liu, 2010b), the National Science Library (Huang, 2010), and the China Institute of Scientific and Technical Information (Bai, 2010). On August 23, 2010, the Shanghai Putuo District Library held a "2010 Frontier Technology Library Forum: Linked Data and the Future of Bibliographic Data" thematic session. Now it seems that linked data has not attracted enough attention in domestic database fields, and there is not yet an influential or mature application in linked data; it is still in the initial stage of exploration.

## 3 Opencsdb - A Scientific Data Linked Network

### *3.1 Linked data applicable to scientific databases*

Linked data has developed four basic guidelines on objects (Burners-Lee, 2009):

1) Use URIs as names for things;
2) Use HTTP URIs so that people can look up those names;
3) When someone looks up a name, provide useful information;
4) Include links to other URIs so that they can discover more things.

In the implementation process of linked data, rule 3 often provides RDF descriptions for resources while the URIs in rule 4 reflect through an RDF link. Based on these criteria, we can conclude that some features in linked data are applicable to a scientific database:

(1) Description - by using RDF, linked data enhances support for different forms of scientific data, including metadata data, data records, and data values of files, becoming semantic. In addition, linked data advocates releasing RDF links between data, similar to hyperlinks between pages. This description mechanism ensures the integrity of the scientific data, which can give full play to the potential

benefits of combining scientific data. By use of an RDF link, a variety of fully autonomous "data islands" are connected together with each other to form a comprehensive repository that provides a rich data source to upper data applications (such as retrieving data integration and data fusion).

(2) Implementation cost - linked data is wholly based on the current web system, which means almost zero upgrade cost to scientific databases. The CSDB project in the "Tenth Five-Year" and the "Eleventh Five-Year" has built a very good web environment, including a domain name system, web servers, application servers, and so on. In the era of linked data, these environments can be used continually. Data publishers will convert the original HTML to publishing data (in RDF format). In this process, in addition to the tools that web site publishers are familiar with, such as FrontPage or Dreamweaver, data publishers will also need the assistance of data publishing tools such as D2R, Pubby, and Triplify.

(3) Specific embodiments - the mechanisms of linked data dispel three concerns of scientific data owners. First, linked data is more of a release mechanism that defines an intermediate format, which is different from the physical storage of raw data. Thus, this feature dispels concerns about data loss or the movement from data collection to data processing that might be disrupted by the surroundings. Second, its correlation mechanism avoids complex data rights disputes, which is conducive to the development of scientific data and prosperity of information. For example, if we extract the chemical composition of a specific plant and then develop synthetic drugs from the extracted composition, the process will involve different databases, such as a plant database, an organic chemistry database, and a drug database. However, using linked data, these databases are connected only by links while the contents still belong to the different data owners. Third, linked data adheres to the web's AAA (Anyone can say Anything Anywhere) concept, which advocates that everyone publishes his/her own data and encourages adding links between external resources (as WWW users add other linked pages to their own home page). In this open environment, no publisher is forced to adopt a centralized data storage center or a unified data expression model.

Through the above analysis, we can see that the concept of linked data is fully applicable to scientific databases and can be better implemented in the CSDB project. Similar to the flourishing development of the World Wide Web, once a good atmosphere is formed for open data, scientific data owners, no matter the scale, will follow the example. This open autonomous mechanism built on a local independent environment is exactly conducive to the sharing of scientific data. OpenCSDB came into being in this context, proposing the concept of building a scientific linked data network with all database resources.

### *3.2 OpenCSDB Construction Standards*

In comparison with the basic rules of linked data, the construction of OpenCSDB follows the following guidelines:

1) In OpenCSDB, each database has a unique URI;
2) In OpenCSDB, every piece of data (data record or data file) has a unique URI;
3) When accessing the above HTTP URI, the data server returns HTML pages or RDF descriptions of data based on the request header parameter values of "Accept" (content negotiation mechanisms of linked data) (Heath, Hausenblas, Bizer, et al., 2008);
4) Each database will publish the RDF vocabularies used so that the consumer can understand the data format (Berrueta, 2012);
5) In addition to the properties of the data themselves, the semantic differences between different data will also be revealed;
6) Each database will open a SPARQL query interface for each database resources description.

As an example, **Figure 1** shows the RDF representation of Qinghai Lake database records about a particular lake investigation (principle 3). The subject "http://qinghailake.csdb.cn/wod/resource/record/investigation/889–2" is the HTTP URI in compliance with data records naming conventions (principle 2); the predicate "investigation: location", "investigation: object" and so on are defined by the RDF vocabularies of the Qinghai Lake database (principle 4).

Principle 5 requires revealing the associations among the data. **Figure 2** shows a more complex example of this, the association between lake survey records, survey objects, and investigators. It can be seen through the link that the survey object points to an animal in the database, "bar-headed goose" (principle 2), and investigators points to a person in the library, "Huang zhiyi" (principle 2).

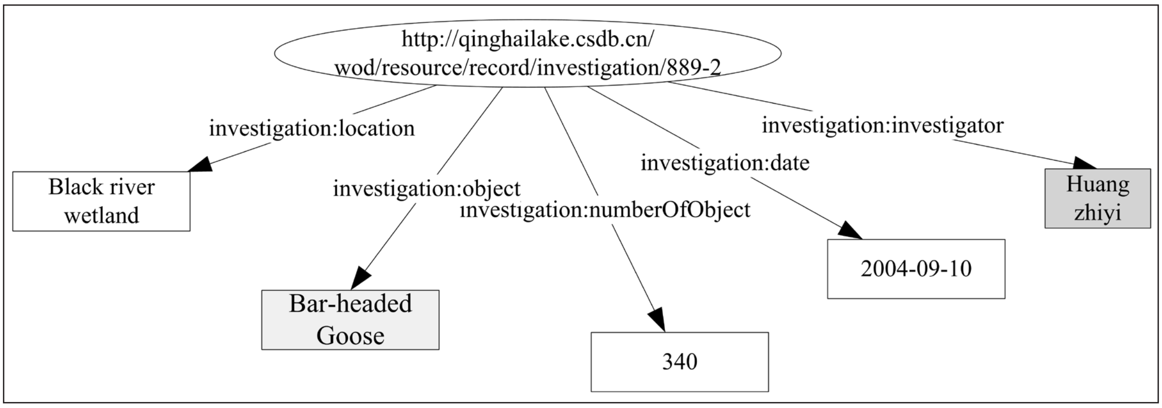
http://qinghailake.csdb.cn/
wod/resource/record/investigation/889-2
investigation:location
investigation:investigator
investigation:date
investigation:object
investigation:numberOfObject
Black river wetland
Huang zhiyi
2004-09-10
Bar-headed Goose
340

**Figure 1:** RDF description of a record in a scientific linked data network.

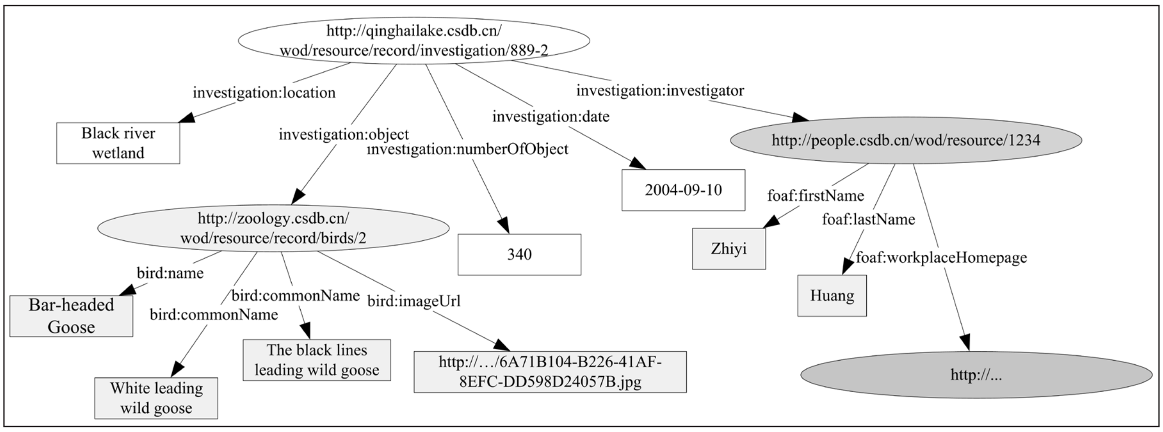

**Figure 2:** RDF description of linked data in a scientific linked data network.

OpenCSDB defines the URI naming principles for data records and data files. The URI design for data records is as follows:

<*baseURI*>/resource/record/<*recordId*>

In this design: *baseURI* represents the URI of the service database site (eg, http://qinghailake.csdb.cn/wod); *recordId* represents the full ID for the record, which is made up of <tableName> / <itemIdValue>; tableName represents the table where the data records are from; itemIdValue represents the primary key value for the record.

Correspondingly, for each data file, the URI design for the description is as follows:

<*baseURI*>/resource/file/<*repositoryName*>/<*fileId*>

In this design: *repositoryName* represents the name of storage location and *fileId* is the unique ID for each file.

OpenCSDB advocates publishing RDF vocabulary that describes the data model (principle 4). The RDF vocabulary of a database contains the following namespace URIs:

<*baseURI*>/vocab/

As for the vocabulary design, OpenCSDB advocates the use of the popular RDF vocabulary on the web (class names, attribute names, etc.), for example, the use of FOAF and vcard vocabulary to define personnel information, the use of dc: title to describe the species name, etc.

## 4 OpenCSDB Software Framework

According to different functionalities and users, the OpenCSDB software framework includes three levels: Building layer, Management layer, and Application layer, shown in **Figure 3**.

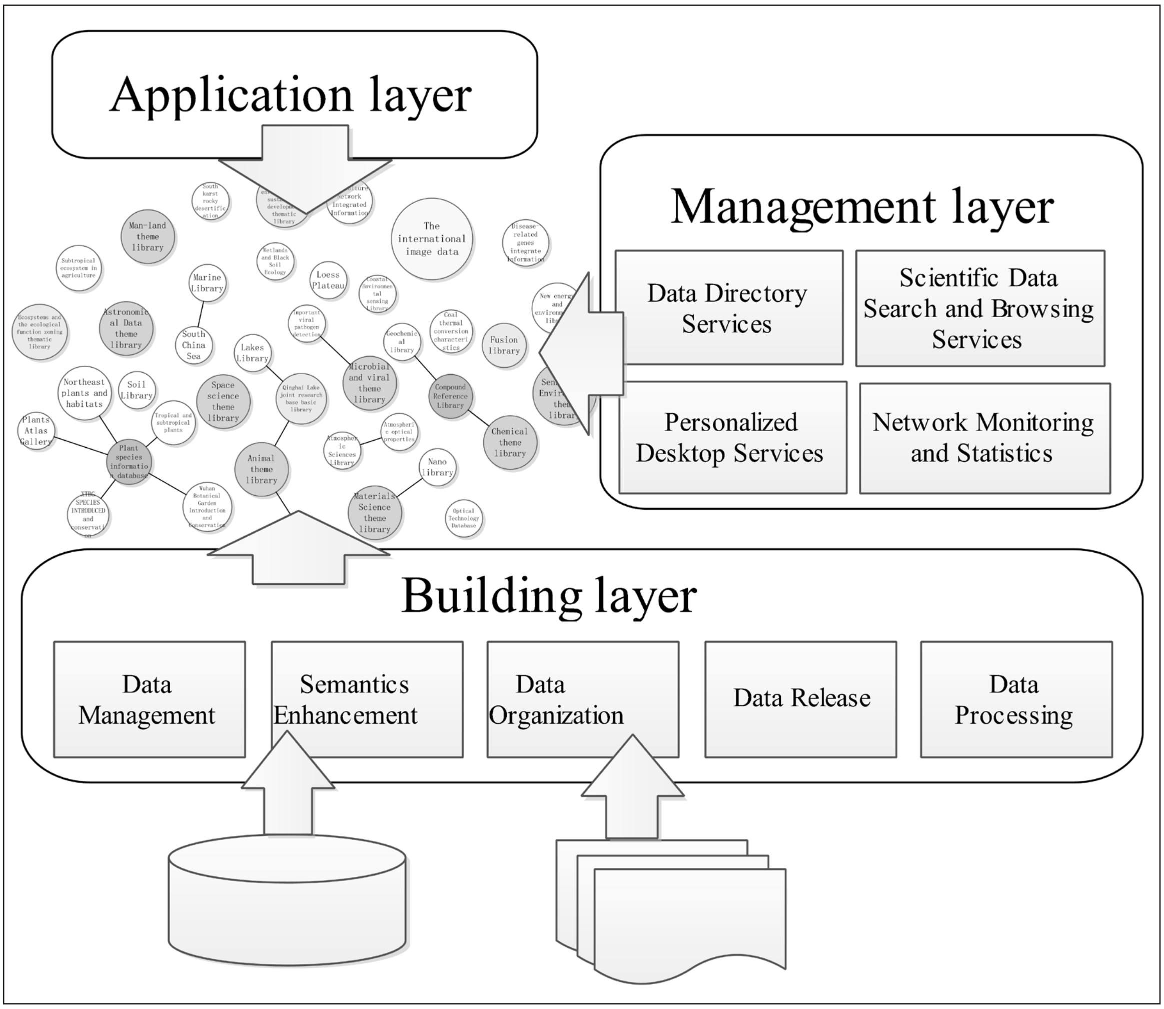
Application layer
Management layer
Data Directory Services
Scientific Data Search and Browsing Services
Personalized Desktop Services
Network Monitoring and Statistics
Man-land theme library
South karst rocky desertification
sustainable development thematic library
Culture Network Integrated Information
The international image data
Disease-related genes integrate information
Subtropical ecosystem in agriculture
Marine Library
Wetlands and Black Soil Ecology
Loess Plateau
Coastal Environmental sensing Library
Ecosystems and the ecological function zoning thematic library
Astronomical Data theme library
South China Sea
Important viral pathogen detection
Geochemical library
Coal thermal conversion characteristics
Fusion library
New energy and environment library
Lakes Library
Northeast plants and habitats
Soil Library
Space science theme library
Qinghai Lake joint research base basic library
Microbial and viral theme library
Compound Reference Library
Plants Atlas Gallery
Tropical and subtropical plants
Atmospheric optical properties
Chemical theme library
Plant species information database
Animal theme library
Atmospheric Sciences Library
Nano library
XTBG SPECIES INTRODUCED and conservation
Wuhan Botanical Garden Introduction and Conservation
Materials Science theme library
Optical Technology Database
Building layer
Data Management
Semantics Enhancement
Data Organization
Data Release
Data Processing

**Figure 3:** OpenCSDB software framework.

The OpenCSDB building layer is used to complete the construction of the data network; its main features include:

1) Data Management: To complete the structured description of unstructured data and localization management of various types of data (data records, data files);
2) Semantics Enhancement: Including the metadata extraction of file data, metadata entry based on templates, database descriptions, and data labels (tagging), and so on;
3) Data Organization: The database classification and catalog organization, URI design and vocabulary design;
4) Data Release: the D2R online map (the E-R model is mapped to the RDF data model), static release of data, 'RDFying' property value (e.g., chemical structure), rule-based link construction;
5) Data Processing: the construction of a theme database from an individual database, link discovery across databases, links between scientific data and scientific literature, and so on.

In addition to software tools, platforms, and the middle layer, the OpenCSDB building layer also includes some norms that are related to the core metadata standards about the database, the URI labeling mechanism of the database, and the data and data standards of the discipline databases.

The OpenCSDB management layer completes the data network management and service. Services provided include:

1) Data Directory Services: Establishing an online catalog for scientific data, providing scientific databases, a registration and navigation directory service about the database, providing registration services about RDF vocabulary, etc.

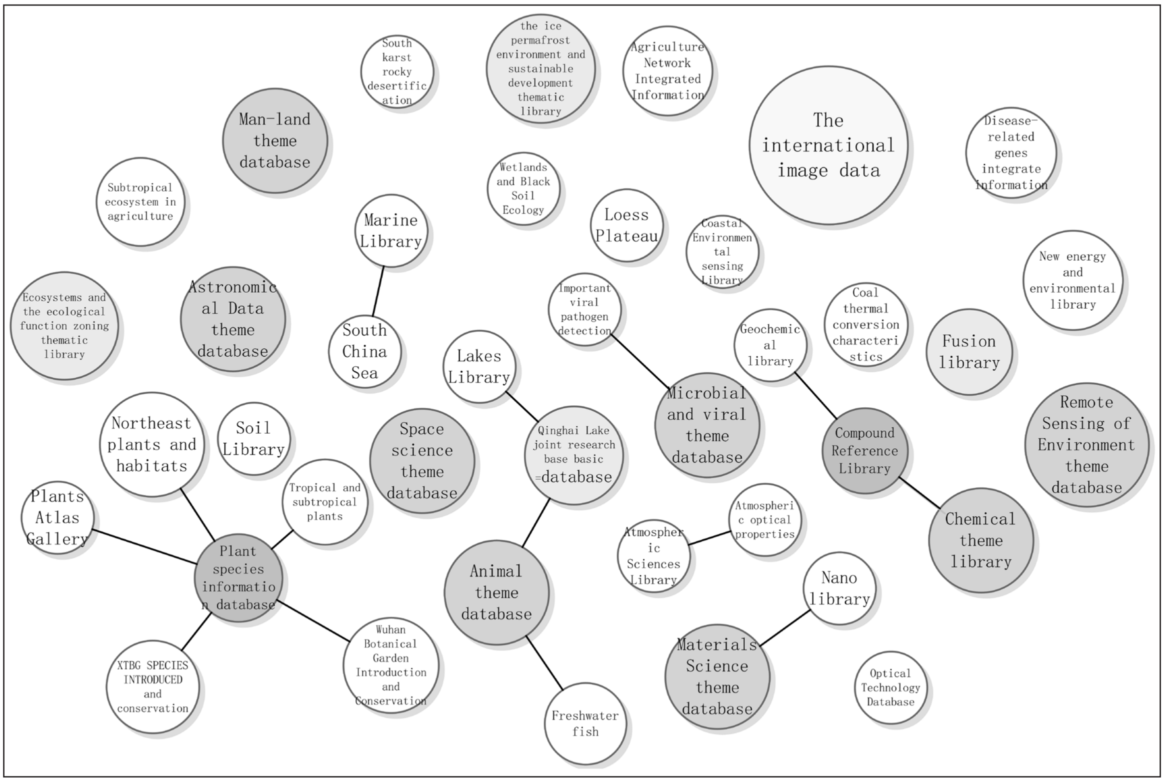
South karst rocky desertification
the ice permafrost environment and sustainable development thematic library
Agriculture Network Integrated Information
The international image data
Disease-related genes integrate information
Man-land theme database
Subtropical ecosystem in agriculture
Wetlands and Black Soil Ecology
Marine Library
Loess Plateau
Coastal Environmental sensing Library
New energy and environmental library
Ecosystems and the ecological function zoning thematic library
Astronomical Data theme database
South China Sea
Important viral pathogen detection
Geochemical library
Coal thermal conversion characteristics
Fusion library
Lakes Library
Northeast plants and habitats
Soil Library
Space science theme database
Qinghai Lake joint research base basic =database
Microbial and viral theme database
Compound Reference Library
Remote Sensing of Environment theme database
Plants Atlas Gallery
Tropical and subtropical plants
Atmospheric optical properties
Atmospheric Sciences Library
Chemical theme library
Plant species information database
Animal theme database
Nano library
XTBG SPECIES INTRODUCED and conservation
Wuhan Botanical Garden Introduction and Conservation
Materials Science theme database
Optical Technology Database
Freshwater fish

**Figure 4:** Scientific data associative network database cloud.

2) Scientific Data Search and Browsing Services: Providing OpenCSDB search engines, providing data resources browsers (based on B / S, or based on C / S architecture), etc.
3) Personalized Desktop Services: Data reference management similar to EndNote, scientific data evaluation and comments, etc.
4) Network Monitoring and Statistics: Providing scientific data resources statistics, providing statistics about data access and updates, the assessment of the scientific data, and the visualization of the scientific database organization.

Cloud is a way to organize the visualization of the scientific database in the management layer. Referring to the LOD cloud, see LOD Data Set Cloud Diagram (2012) (http://richard.cyganiak.de/2007/10/lod/lod-datasets_2010-09-22.html), the "Eleventh Five-Year" scientific database cloud (http://standards.csdb.cn) is formed as shown in **Figure 4**.

In OpenCSDB cloud, each circle represents a database, and the line represents the association between them (it can be seen in the cloud that the associations are centralized around reference databases). The color, position, and radius of each circle have their special meanings. For example, the radius of the circle corresponds to the number of data items in the database and the color of the circle corresponds to the disciplines of the database.

That OpenCSDB application layer, which is uppermost, is used to develop specific e-Science applications for specific needs. In this layer, OpenCSDB provides the access interface specification for data and improved APIs.

## 5 The Application of OpenCSDB

Late in the "Eleventh Five-Year" in the CSDB project, a prototype of the OpenCSDB software framework was established, and some functions were implemented.

As for the norms, the CSDB project has established a "Scientific Database Core Metadata Standard (Version 2.0)" (http://standards.csdb.cn) and a "professional database, theme database, reference database" scientific organization system, developed programmatic access interfaces for individual databases, completed URI designs for 528 databases and sub-databases, and developed URI naming conventions for scientific database records and data files.

As for software, embedding VDB-WOD middleware on the basis of visual data management and publishing tool VisualDB2.0 can realize data publishing in RDF for data records and data files. The latest version of

VisualDB has added semantic enhancements, including metadata extraction, metadata tagging, classification system importing, data catalog establishment, and so on. Based on the VDB-WOD interface, a scientific data search engine named Voovle (http://voovle.csdb.cn) was developed to provide keyword-based fuzzy searches and exact match searches. What is more, based on the link discovery tool Voovle-LDT, we can generate new cross-database links with the specified rules. Scientific data resources and the Services Registration System RSR, Resources and Services Registry, CSDB (2011) (http://rsr.csdb.cn) were developed to complete online registration and collection of core database metadata, and they can be seamlessly integrated into Voovle. A statistical system about the online resource visiting situation RESSTAT was developed that completes the monitoring statistics of the VDB-WOD service interface, compiles online statistics of the amount of resources and data stored, implements a comparison between the number of data records and data files, and gives visual statistics reports.

As for the application effects, during the "Eleventh Five-Year", by using VisualDB, OpenCSDB collected a total of 52 scientific databases from more than 40 Chinese Academy of Sciences Institutes. RSR provides metadata injection services for 528 databases and sub-databases; Voovle provides a semantic search service involving 124 databases and 5.64 million pieces of scientific data from 37 database construction institutes. RESSTAT also provides an online resources statistical service for nearly 150TB of databases.

## 6 New Challenges and Prospects

Due to the quality of scientific databases and the limitations of linked data, OpenCSDB encountered new challenges in the implementation process, including the following:

(1) A semantic consistency problem that linked data exposed within the data. Heterogeneous phenomena from different sources and different institutions are prevalent in scientific databases. This has seriously affected scientific data sharing (Chen, 2012). As the existing scientific databases were built separately by each institute, they lack appropriate norms and standards and show complexity in knowledge representation in specific fields. As a result, significant difficulties occur for the specialization of metadata standards and their reference models in a specific field. Coupled with the lack of mature technical specifications and tools, data managers have difficulty in establishing a mature conceptual model of a specific field. Therefore, the fact that each institution builds its own scientific databases, using different identification systems, different classification systems, different vocabularies, and data models results in homonyms, synonyms, and synonymous and heterogeneous problems that bring greater obstacles to the semantic consistency expression of scientific data.

(2) There is a conflict between publishing scientific data statically and processing scientific data dynamically. The generation, transmission, processing, and application of scientific data is an ongoing, continuously iterative, changing process, and the associated context (such as processing procedures) is also in a constant state of change. On the contrary, the resource description of associated data is often considered to be a static section (a description of data should not be changed frequently otherwise the application will crash easily). This dichotomy exposes the limitations linked data has in scientific data sharing. Sean Bechhofer et al. (2011) presented the 7-Re standard in the scientific data environment, which stands for Reusable, Repurposeable, Repeatable, Reproducible, Replayable, Referenceable, and Revealable. To achieve this standard, the CSDB project requires more effort.

(3) Data access control problems. Linked data has not made more standardization work in user authentication and data access control (at least there is not yet a standard method that can be recommended), and therefore the application server must implement unification authentication and access control in the process of data access, which limits the interoperability between different systems. Although OpenCSDB has proposed six principles for data network construction, SPARQL services only open to the specified URI. How to apply a standardized mechanism for a linked data set to achieve access control to databases, classes, attributes, and so on is an urgent problem.

(4) Search and sort problems for massive data. OpenCSDB contains vast amounts of scientific data. Because of the unique characteristics of scientific data such as diverse formats, rich attributes, and being structured and interrelated, traditional search engines have been unable to meet OpenCSDB requirements. Therefore search engine technologies based on linked data networks should be studied to complete capture of massive data through sitemap.xml or SPARQL protocol. Furthermore, the ranking mechanism of a scientific data document should be studied to optimize the index and retrieval for scientific data in order to achieve optimized scientific data search services.

Overall, despite all these difficulties and challenges, linked data is still regarded as a standardized, practical, open access mechanism. Unlike in the previous web of documents, using the linked data mechanism, we can establish and improve the scientific data network during the "the Twelfth Five-Year ". On the web, each database and each piece of data can be open and accessible, and the semantic content will be returned. Throughout linked data applications (Hausenblas, 2009), as well as in the "scientific data popularization" (data publishing, data references, data traceability, etc.) in the book publishing industry in recent years (De Schutter, 2010), it is confidently believed that OpenCSDB will play a more important role in scientific data sharing with its further promotion and application.

## 7 Acknowledgement

The work is supported by the Director Foundation of the Computer Network Information Center of the Chinese Academy of Sciences (Granted No. CNIC_ZR_201304; CNIC_PY_1405).

## 8 References

Bai, H. (2010) Ordering Deep of Information Organization based on Linked Data. *Library Frontier Technology Forum 2010*, Shanghai, pp 2010 - 2018.

Basic Database of Joint Research Center of Chinese Academy of Sciences and Qinghai Lake National Nature Reserve (2012) Retrieved from the World Wide Web December 8, 2014: http://www.qinghailake.csdb.cn

Bechhofer, S., Buchan, I., De Roure, D., et al. (2011) Why linked data is not enough for scientists. *Future Generation Computer Systems*.

Berners-Lee, T. (2006) Linked data - Design Issues. Retrieved from the World Wide Web December 8, 2014: http://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T. (2006) Linked data FAQ. Retrieved from the World Wide Web December 8, 2014: http://structureddynamics.com/linked_data.html

Berrueta, D. (2012) Best Practice Recipes for Publishing RDF Vocabularies. Retrieved from the World Wide Web December 8, 2014: http://www.w3.org/TR/swbp-vocab-pub/

Chen, W. (2012) On the Individual Identification and Interdisciplinary Integration for the Scientific Data. *Proceedings of Scientific Database and Information Technology*.

Day, M. (2005) *DCC Digital Curation Manual: Installment on Metadata*.

De Schutter, E. (2010) Data Publishing and Scientific Journals: The Future of the Scientific Paper in a World of Shared Data. *Neuroinformatics 8*(3), pp 151–153.

Diseasome | Map: explore the human disease network. Dataset, interactive map and printable poster of gene-disease relationships. Retrieved from the World Wide Web December 8, 2014: http://diseasome.eu/map.html

Gray, J., Liu, D.T., Nieto-Santisteban, M., et al. (2005) Scientific data management in the coming decade. *ACM SIGMOD Record 34*(4), pp 34–41.

Hausenblas, M. (2009) Linked Data Applications. *First Community Draft, DERI*.

Heath, T., Hausenblas, M., Bizer, C., et al. (2008) How to publish linked data on the web. In *Proceedings of LDOW 2008.*

Huang, Y. (2010) Research on Linked Data-Driven Web Applications. *Library Journal*.

Jonquet, C., LePendu, P., Falconer, S., et al. (2011) NCBO Resource Index: Ontology-based search and mining of biomedical resources. *Web Semantics: Science, Services and Agents on the World Wide Web*.

Liu, W. (2010) Linked data: meaning and implementation. http://www.kevenlw.name/?p=1435

Liu, W. (2010) The Web of Data. http://www.kevenlw.name/?p=1185

LOD Data Set Cloud Diagram (2012) Retrieved from the World Wide Web December 8, 2014: http://richard.cyganiak.de/2007/10/lod/lod-datasets_2010-09-22.html

Momtchev, V., Peychev, D., Primov, T., et al. (2009) Expanding the pathway and interaction knowledge in linked life data. In *Proceedings of International Semantic Web Challenge*.

Open Data Protocol (OData) (2011) Retrieved from the World Wide Web December 8, 2014: http://www.odata.org/

Resources and Services Registry, CSDB (2010) Retrieved from the World Wide Web December 8, 2014: http://rsr.csdb.cn

Resource Statistics System, CSDB (2011) Retrieved from the World Wide Web December 8, 2014: http://resstat.csdb.cn/

Scientific Data Formats (2012) https://nf.apac.edu.au/facilities/software/IDL/docs-6.2/sdf.pdf

Scientific Database Core Metadata Standard (Version 2.0) (2004) Retrieved from the World Wide Web December 8, 2014: http://standards.csdb.cn

Scientific Database Platform, Chinese Academy of Sciences (2011) Retrieved from the World Wide Web December 8, 2014: http://www.csdb.cn

Shen, Z., Li, J., Li, C., et al. (2011) VisualDB: Managing and Publishing Scientific Data on the Web. *International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*.

SweoIG/TaskForces/CommunityProjects/LinkingOpenData (2010) Retrieved from the World Wide Web December 8, 2014: http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

Voovle (2010) Retrieved December 10, 2014 from the World Wide Web: http://voovle.csdb.cn

W3C Library Linked Data Incubator Group (2005) Retrieved from the World Wide Web December 8, 2014: http://www.w3.org/2005/Incubator/lld/

**How to cite this article**: Shen, Z, Li, J and Han, F 2015 Opencsdb: Research on The Application of Linked Data in Scientific Databases. *Data Science Journal,* 14: 4, pp. 1-11, DOI: http://dx.doi.org/10.5334/dsj-2015-004

**Published:** 22 May 2015