A genealogy data services platform implemented with linked data technologies①

XIA Cuijuan[1*], LIU Wei[1], CHEN Tao[2] & ZHANG Lei[1]

1 Shanghai Library, Shanghai 200031, China

2 Shanghai Information Center for Life Science, Shanghai 200031, China

Abstract

Open data becomes not only the responsibility but also a development opportunity for the government, scientific research institutions, libraries and other cultural heritage institutions. The Shanghai Library, starting from the genealogical data, has been dedicated to constructing the historical documents and data services platform through reorganizing its traditional resources by utilizing the Linked Open Data technologies. The genealogy data services platform, through BIBFRAME-based ontology design, data transformation from RDB to RDF, system design on the basis of the four principles of Linked Data as well as the system development based on the framework of semantic technologies, supports bibliographic control in the Internet environment and satisfies general users' needs for tracing their family roots, and professional researchers' demands for data mining.

Keywords

Genealogy, Data service, Linked data, Open data

0 Introduction

Open data has been a new trend during the Internet development. Data as a vital resource has reached a consensus worldwide. Governments and public institutions, holding the most public data, have stood at the vanguard of the open data movement (Wu, 2015). In 2009, Data.gov officially launched in the United States, driving the data open movement, and followed by Data. gov.au of Australia and Data. gov.uk of the United Kingdom. In November 2010, the "EU Commission's Open Data Strategy" was first proposed by the European Commission and put the data open movement to its heyday (Attard, Orlandi, Scerri, & Auer, 2015). Media including *The New York Times*, and the BBC successfully implemented the strategy and the library community was the active advocate of the data open movement. National libraries of Sweden, the United States, Hungary, the United Kingdom, Germany, Spain, South Korea, Japan and other countries around the globe as well as OCLC published their bibliographic data or standard data in the form of linked data. The Library of Congress also took the lead to carry out the liked data of

① This article is an outcome of the youth project "The Application of W3C's RDB2RDF Standards in Building Linked Data Services" (No.13CTQ008) supported by National Social Science Foundation of China.

* Correspondence should be addressed to XIA Cuijuan, Email: cjxia@libnet.sh.cn, ORCID: 0000-0002-1859-6979

bibliographic data format standard.

The Shanghai Library has been paying close attention to open data movement and started tracking, researching, and developing relevant technologies from early on. Open data movement is deemed as a new opportunity to bring the digital library into the next generation internet which is featured by the data technology. For the Shanghai Library, although the digitization of the paper documents as well as the literature retrieval service of the large amount of historical literature resources such as ancient books, genealogies, correspondences, modern documents, documents of the Republican period of China, archive, photographs, notes, manuscripts, and tabloids has been ongoing, to better satisfy the readers' demands, the knowledge contained needs to be described. By using new technologies to provide services on the internet to more users, the open data technologies involve in the optimization and iteration of system as well as the construction of the contents so as to realize the platform system, which will become an indispensable place for readers to engage themselves in their study, communication and research.

Chinese genealogy is one of the most important characteristic literatures of the Shanghai Library. After long-term research and arrangement, the Shanghai Library has made some influential achievements, including compiling and publishing *Summary of Genealogies Held by Shanghai Library, The General Catalog of Chinese Genealogies, General Theory of Chinese Genealogies, Selected Documents of Chinese Genealogies. The General Catalog of Chinese Genealogies*,which collects more than 54,000 kinds of genealogies from over 600 institutions in Japan, Korea, North America, Germany, Australia, and other regions and countries, includes 608 surnames and separates out over 70,000 ancestors and celebrities, 1,600 places, more than 30,000 clan temple titles. The catalog is not only a union catalog of Chinese genealogies but also an encyclopedia of Chinese genealogy knowledge. Although these precious achievements exist in paper or image forms and the content and index revelation is only for publication purpose or provides simple search fields, they provide a foundation for developing a knowledge services platform on the basis of linked data.

Through many years of research and exploration, the technical research and development team of the Shanghai Library expects that it is the time to apply the new data management technology represented by linked data. The technologies will enable libraries to make full use of the long-term accumulated literature research achievements and to make fine-grained descriptions of the data, facts and other knowledge within them. Coding methods and technical means of network knowledge organization are utilized to reorganize the resources. The Internet platform in web-scaleis used to achieve the purpose of bibliographic control. As for the genealogical data, they need to support various services such as faceted visual browsing, semantic search and knowledge mining for researchers in humanities while satisfying general users' needs for tracing their family roots. The linked data technologies contribute to breaking the traditionally isolated condition for different kinds of resources, advancing data open, promoting knowledge flow and giving full play to the potential value of the resources in the process of open use.

1 Functional requirements

1.1 Library bibliographic control requirements

Libraries take on particular function and mission for bibliographic control. *The General Catalog of Chinese Genealogies* published in 2005 was presided over by the Shanghai Library and compiled by numerous genealogy research scholars. The investigation, textual research, and compilation of collection distribution of Chinese genealogy documents were carried out around the world. The catalog contains abundant and detailed literature description and content description information. The requirements of bibliographic control in web-scale and providing services at any time and any place can be realized if the new technology under network environment is developed and utilized.

To be specific, key requirements of bibliographic control for genealogy catalog are embodied in the following three aspects:

(1) Establishing global genealogy union catalog, and promoting the reuse and sharing of data. One of the important functions of bibliographic control is to clarify the collection condition of a certain resource in different institutions and provide clues on how to access the resource. With the development of internet technology, it becomes possible to establish a national and even global genealogy union catalog. *The General Catalog of Chinese Genealogies* has already provided a very solid data foundation. The first step to constructing genealogy services platform is to import the existing data and provide readers with information such as the existing editions of a certain genealogy, its collection condition around the globe and the approach to access the document. At the same time, the platform should facilitate the reuse and sharing of the knowledge created by organizations and genealogy research experts among different institutions.

(2) Conducting Web-based authority control. Authority control is a critical part in bibliographic control. The essence of authority control is to realize the concept-based description and matching (Liu, Zhang, & Xia, 2015). It is necessary to differentiate different kinds of standard entities such as person, institution, place, event from the conceptual level and represent the concept by using unified text labels. For genealogy resources, to resolve problems like identification, disambiguation, and merging of different expressions of the same names of persons, places, and dynasties, the models of concepts including surname, people, institution, place name, dynasty in Chinese historical calendar need to be established and the relationships among these concepts need to be expressed through specific semantics. On the basis of Web-based authority control, the controlled authority term is required to be uniquely identified and located in web-scale, and the semantic description information of it should be acquired, recognized and understood by machines.

(3) Supporting the sustainable development of bibliographic control. The printed version of *The General Catalog of Chinese Genealogies* is a static and closed document and its bibliographic

data were up to 2003. Therefore, we need to establish an open platform and further supplement and optimize the existing bibliographic records such as the personal information of one's ancestor or celebrity and the place information of family migration in certain kinds of genealogy. Most importantly, new bibliographical records including newly-collected ones and the records about to be collected should be easily added.

1.2 Differentiated user needs

To construct a genealogy services platform, we have to put users' needs in the first place. Since the Shanghai Library opened the first genealogy reading room since 1996, we have accumulated abundant experience in user services. According to users' different purposes for utilizing genealogy resources, the user needs can be divided into the following levels:

(1) Seeking family roots and ancestors based on given information. For general users who need to trace their family roots and ancestors, they usually search for related genealogy documents, relevant ancestors or celebrities, relatives' personal information (such as dates of birth and death, or life events) based on given information like the name of one's ancestor or celebrity, family location, or clan temple title. This type of users requires not only the documents themselves but also the contents such as data, facts and knowledge included. These contents, for example, contain the detailed information of the ancestor or celebrity or relative in one's family, the kinship between the characters in different genealogy documents, and family migration route. Although the user need at this level is not complicated, it has a relatively high demand for the precision ratio. The existing keyword matching search leads to noises to a certain extent. The search needs to be based on the matching between concepts in order to locate exactly the results that the readers are looking for. The genealogy services platform should not only offer the convenient way for users to access documents but also provide them directly with the contents they want.

(2) Knowledge discovery oriented to a specific research subject. For researchers in humanities, genealogy is one of the vital research materials apart from official historical records and chorographies. The unique value of the genealogy has been widely acknowledged by academics. However, the precious intellectual legacy is hidden behind the vast and numerous volumes of genealogy. If only depending on the current system which organizes resources by documents and is based on simple fields and keywords for search, it is labor intensive and time-consuming to explore the data, facts, and knowledge which are scattered among the hundreds of thousands of volumes of genealogy materials. Thus, it is particularly important to develop the user-friendly knowledge navigation as well as knowledge discovery functions.For example, by clustering display on the surname, place, clan temple title and dynasty and discovery for the associated relationship between concepts and entities, such knowledge as the relationships between people and places, and interpersonal relations, the geographic migration route of a family, the distribution of a surname

in one region, and the distributed living places of a clan in different genealogy documents can be found.

(3) Knowledge evolution and accumulation based on User Generated Content (UGC). Genealogy is the historical land chart and census register which records the stemma and its stories or the development process of the family reproduction by the consanguineous group of the same clan and ancestor. There are a large number of nongovernmental groups and communities who have a deep understanding and research on their own surnames and genealogies. They are both the library users and experts on genealogy study. They have more comprehensive and deeper understanding on certain surnames or certain family genealogies than librarians have. If we construct an open platform which can not only provide the genealogy materials collected by the library, but also facilitate positive interaction and communication between users, users and collection institutions, while reorganizing, processing and preserving the knowledge generated from communication, the purpose of optimizing the genealogy knowledge base as well as augmenting its value in communication and dissemination can be achieved. Along with the popularity of Web 2.0 technology and the widespread of UGC's concepts of "Crowdsourcing" and "Crowdfunding", it is common to take users' behaviors into the opreation process of the library (Fan, 2011). As a result, the new genealogy knowledge services platform should not only be a static featured database but also the organism supporting the continuous growth and evolution of knowledge.

2 Design implementation

Linked data technology is implemented in building the genealogy knowledge services platform because it organizes knowledge based on domain conceptual system (ontology) instead of documents, and describes and retrieves knowledge via RDF universal data model which can be written in an RD triple by the grammatical constitutes (subject, verb, object). By using the developed data proof and knowledge mining tools to support the maintenance and update of knowledge, users are allowed to access parts of the document data instead of the whole document. On the other hand, with the widespread and profound application of linked data in libraries (Mitchell, 2013), a set of technology, methodology and process of metadata, ontology, RDF data conversion, RDF data storage and querying, and data visualization has been formed and can meet the requirements of bibliographic control and authority control, data reusing and sharing, knowledge organization and discovery.

The design of genealogy knowledge service platform by the Shanghai Library went through the following process. Firstly, we designed the genealogy ontology which was downward compatible, easy to extend, convenient for reusing and sharing, and supports the data reorganization and knowledge modeling of genealogy. In addition, concepts such as person, institution, place, event and their relationships in genealogy were explicitly defined. Secondly, we cleaned the existing genealogy metadata and extracted conceptual entities and assigned HTTP URI for each entity. Based on the

RDF abstract data model, we described the entities and the relationships between these entities, and associated them with external data so as to enrich the data semantics if necessary. The data, after encoding by the machine-readable RDF serialized format, were stored in the dedicated RDF database. Finally, we designed the system based on the four principles of linked data published, developed framework access operation on data by semantic technologies, utilized visualization technologies to present data, and made use of Web 2.0 technology to support users in contributing their knowledge in order to achieve functions of knowledge navigation, discovery and evolution.

2.1 Design of linked data model based on ontology

Ontology is the sharable and reusable conceptual model being formed after the domain knowledge is abstracted. It is usually represented by a systematic terminology and formalized description of the interrelations, and becomes the code system which can be recognized and processed by machines after it is encoded by a certain machine language. Ontology endows data with semantics and is regarded as the container of the knowledge in the data. On the basis of the ontology design principle which emphasizes on the reuse of existing terminology as much as possible, the genealogical ontology by the Shanghai Library is primarily based on the BIBFRAME 2.0 led by the Library of Congress, reuses parts of the terms from FOAF、Geonames、Schema.org vocabulary lists, and customize featured properties of genealogy resources.

Genealogical ontology is designed on the basis of BIBFRAME. On the one hand, BIBFRAME, a replacement for the traditional standard of MARC, is the new standard for bibliographic data formats and can be used together by libraries and other memory organizations such as archives, museums, and museums. The BIBFRAME has very good compatibility, extendibility and openness. The BIBFRAME vocabulary can describe the features of genealogy resources very well. On the other hand, BIBFRAME is designed as a linked bibliographic data model and its core classes of “Work-Instance-Item” are the simplified FRBR (Functional Requirements for Bibliographic Resources)(Library of Congress, 2015). The model can meet the requirements of bibliographic control very well and its data model concepts including person, institution, family and event are suitable for describing entities related to genealogy resources as well as meet the needs of authority control.

Moreover, the genealogical ontology reuses the terminology in FOAF to describe the ancestor or celebrity in genealogy and customizes the characteristic properties of Chinese historical figures such as “genealogy name, courtesy name, pseudonym name, and posthumous name” as supplements. The terms of Geonames are used to describe the places, terms from Schema.org and W3C Organization for genealogy related collection organizations and W3C Time Ontology for describing time information.

Some properties are self-defined and customized to describe the information of dynasties

in Chinese history. In order to facilitate the sharing and reusing of genealogical ontology, the Shanghai Library has released to the public the genealogical ontology on the Web, encoded in RDFs and OWL. What's more, the website supports three kinds of view modes to bring convenience to specialists in this field for a deeper understanding of the ontology. The visualization of Model View displays the relationship between genealogical ontology classes and properties. Class View, through the hierarchical relationship between superior class and subclass, presents the classes and properties. List View demonstrates the classes and properties according to the alphabetic list of the name of classes as well as property. The genealogical ontology and all the RDF data can be downloaded from the website (See Figure 1) .

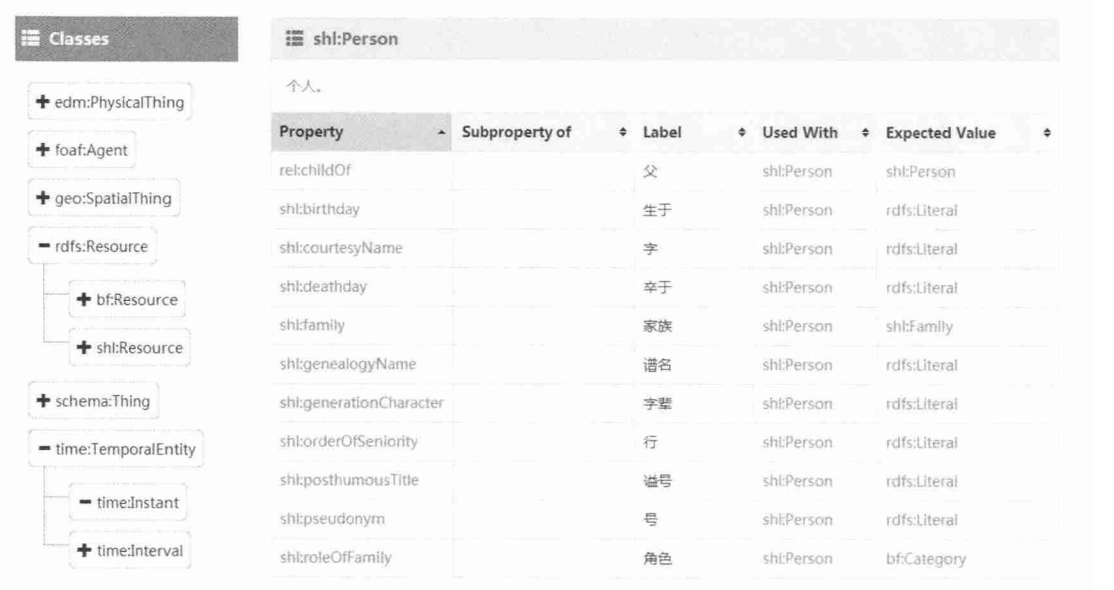

Figure 1. The genealogical ontology of shanghai library published by linked data.

2.2 Standard open data format: From RDB to RDF

The third principle of linked data requires the RDF format of data. The abstract format of RDF data and a variety of serialization formats such as RDF/XML、Turtle、JSON-LD are the standards recommended by W3C. These standard data formats are cross-platform, open and can be processed by different program languages (Cyganiak, Wood, & Lanthaler, 2014).

The RDF data of the genealogy knowledge services platform are generated from the pre-existing metadata not only from *The General Catalog of Chinese Genealogies*, but also from the metadata of the newly added genealogy documents. First of all, entities such as work, instance, item, person, institution, name of place are extracted from metadata and each entity is assigned with a HTTP URI. Then the classes and properties which are defined by genealogical ontology are used

to describe the association between these entities. I listed in detail the correspondence between genealogical metadata and genealogical ontology in my article *A Genealogical Ontology in the Form of BIBFRAME Model*

The data of *The General Catalog of Chinese Genealogies* are stored in Excel tables and the collected genealogy data are stored in SQL Server in the form of MARC. Both of the data can be taken as the RDB data format of "Record-Field-Field Value". Thus, the data format needs to be transformed from RDB to RDF, the process of which is called RDB2RDF. There are two open source conversion tools that can be applied to accomplish RDB2RDF process: one is "DB2Triples" that supports W3C RDB2RDP standard and the other is "OpenRefine". Both of the two tools support the creation of the mapping between the tables & fields in RDB and the classes & properties in genealogical ontology, define the generated standard of URI, and automatically

Figure 2. The mapping and conversion from RDB to RDF.

generate RDF data. There are also some differences between the two tools. The DB2Triples applies R2RML standard of W3C and supports one-time access to multiple relational databases, as well as generates RDF data of multi-class entities. However, the disadvantages of DB2Triples tool lie in that its ontology mapping requires the configuration files of JSON languages to edit text formatting and lacks the user-friendly interface (Xia & Jin, 2015). Although OpenRefine is not convenient for operating the data from multi tables at the same time, it is given the so-called WYSIWYG (What you see is what you get) user interface. Therefore, when transforming the collection genealogy data in multiple relational databases stored in SQL Server, the DB2Reiples tool is applied. When transforming the data of *The General Catalog of Chinese Genealogies* in a single Excel table, the OpenRefine tool is applied. Figure 2 takes the ancestor and celebrity in the genealogical table, for example. It presents the process of transforming the data in Excel table from Turtle format to RDF format by using OpenRefine tool.

2.3 The design based on the four principles of linked data and the implementation based on semantic technologies

The design of system follows the Four Principles of Linked Data. After investigations on Cool URIs standard (Sauermann & Cyganiak, 2008) and other linked data projects conducted by international governmental sectors and libraries by combing with the actual demands, we formulated *The URI Design Specification of Shanghai Library* and generate HTTP URI for various entities in genealogy data according to this standard. With respect to the descriptive information for entities, they are organized by RDF abstract data model and encoded by standard serialized format. When visiting the HTTP URI of an entity, the relevant RDF information about the entity will be obtained. It also supports content negotiation mechanism. When a user visits via ordinary browser, the system returns to the Html page for people to read. However, when the semantic-web browser or semantic proxy (program) is used to visit URI, the system returns to the corresponding formats of RDF data (such as RDF/XML, RDF/Turtle, JSON-LD) according to the requester's request for the content format delivered by Http reader.

The development of the system is based on the semantic technology framework. The RDB2RDF tools supporting W3C's RDB2RDF standards are used during the transformation process of data from RDB or EXCEL format to RDF/Turtle format. And other data cleaning and transformation tools like OpenRenfine are also used. Then the RDF data generated by those tools can be loaded into the RDF store (Open Link Virtuoso) instead of traditional RDB database, and the data interaction between the visualization layer and the storage layer is driven by SPARQL via Jena.The data visualization tools such as SIMILE Timemap, Baidu Echarts and AMAP(高德地图) are used to provide visualized data service to the end users. The data of RDF are stored in the RDF storage called Open Link Virtuoso. Between DF storage ad visualized presentation layer, data can be

queried and accessed by RDF query language SPARDL.Jene, as the development tool, is utilized to process RDF data. A variety of tools such as SIMILE Timemap, Baidu Echarts, AutoNavi map is utilized to realize the visual display of data. All the development framework is shown in Figure 3.

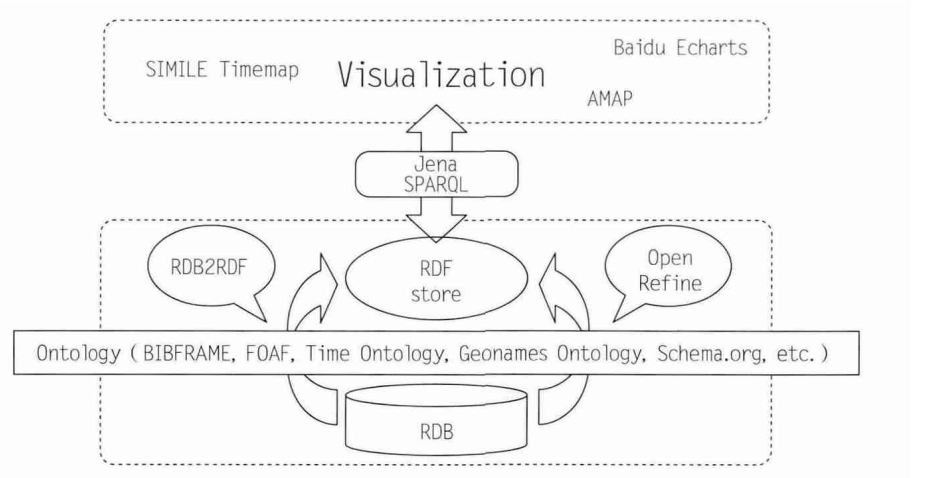

Figure 3. Development framework based on semantic technology.

2.4 System functions designed for bibliographical control , knowledge discovery and knowledge evolution

There are three important functions to be considerate while the platform was designed. Firstly, we need to meet library needs for functions of bibliographic control and data sharing, including the full presentation to the editions, copies, and collection organizations of a certain kind of genealogy and the authority control of a variety of content entities, as well as the data consumption interface used for data reuse and sharing. Secondly, the knowledge discovery function oriented to the general public and researchers in humanities needs to be designed, including retrieval function based on conceptual matching as well as the visualized browse function based on the associated relationships between person, place, institution, and time. Thirdly, the knowledge evolution function should be developed. This function supports users to make modification and supplement on the pre-existing data, and save, organize and process this knowledge.

2.4.1 Bibliographic control

The requirement of genealogy bibliographic control is supported by BIBFRAME core data model (Work-Instance-Item). But in reality, it needs to be further simplified according to the concrete requirements of genealogical description. If strictly practicing the definition of BIBFRAME, the relationship between the entity relational model of work-instance-item is 1: n and 1: n, i.e. 1 work

corresponds to multiple instances, and 1 instance corresponds to multiple items. As a result, taken genealogy as a "work", then the different editions (copies or photocopies of a certain edition) of the same kind of genealogy can be regarded as "instance". Thus, information such as publication date and place can be described every time the genealogy is produced or published. But for the current existing genealogy data, although the publication time and institution of the wooden copy has been retained, the detailed information of the copies and photocopies is not kept; only the collection information is reserved. For example, the five volumes of *The Genealogy of Shang Kuan Family,* their original wooden version printed by TianShui Tang, which are preserved in "Seeking Original Surname" and the GSU (Genealogical Society of Utah) in the United States, and the copies are preserved separately in the Fujian Library and Lin Jiashu sector of Taiwan Affairs Office in Zhangzhou City, Fujian Province. The collection information is the remarkable feature of "item". So we decide to take these copies as the different "items" for the "wooden copy" instance. Consequently, the entity relational model of genealogy's "work-instance-item" is 1:1: n, as shown in Figure 4.

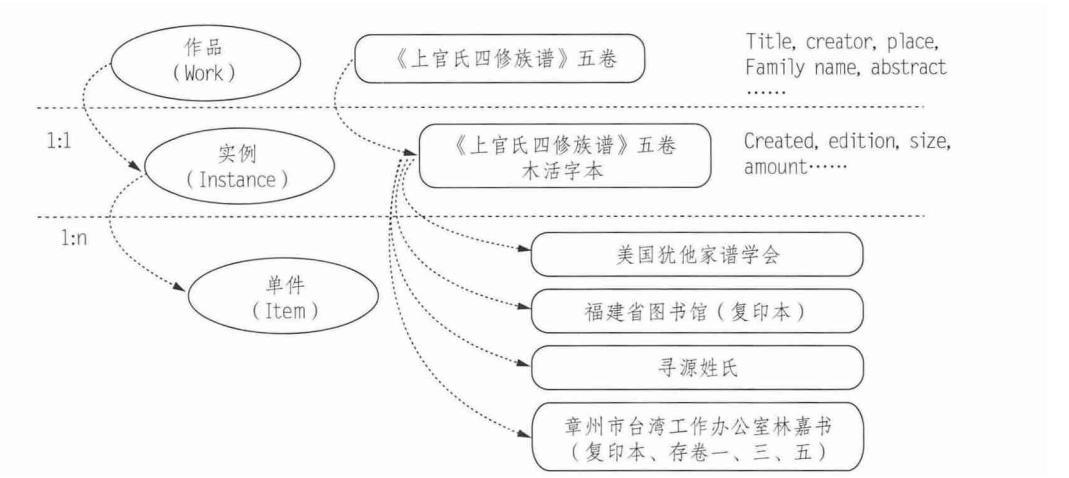

Figure 4. The genealogy bibliographic control model based on BIBFRAME2.O by Shanghai Library.

2.4.2 Knowledge discovery

(1) For the general public

To meet the needs of the general public for browsing at random and seeking ancestors, we took the conciseness, novelty, and interest into consideration when designing the interface and functions. The homepage is designed in the pattern that directly presents relevant genealogy documents and statistics of the ancestor or celebrity according to the surnames so as to attract users to click on. Users can directly keep in touch with the data and knowledge and discover the information. This design will appeal users to continue to explore. Users, therefore, can start from surnames to acquire the knowledge of the origin, evolution as well as the celebrities in history without professional

knowledge, and then gain a sense of identity. At the same time, users can find out the information about the collected ancestors as well as the genealogy documents on this platform. Click on "Collection Information", users can learn further information about all the collection organizations and their location. For example, if this genealogy is collected by the Shanghai Library, a full-text image link will be displayed directly. On the right area of the homepage, there are functions as basic search portal and advance search portal being displayed in the mode of automatic carousel. Users can input retrieval words to conduct the query.

For querying for the documents without a specific goal, the concrete genealogical title and the family's living place name, the space-time diagram can be utilized to explore. Time is designed as a knob. So when users rotate the knob or input the year, the genealogy created in this year will be displayed and the location of the place will be marked with a flag on the map. Click on the flag, all the eligible genealogies will be shown.

(2) For professional researchers

Professional researchers include the genealogy experts specialized in a certain surname and a family, or researchers in other domains who use genealogy documents to provide materials and evidence for their research subjects. For this type of users, the platform develops such functions as conceptual-matching based advanced search as well as temporal-spatial correlation-based discovery function.

Advanced search supports precise search by conceptual entities such as surname, author, the family's living place name, clan temple title, ancestor or celebrity, and collection organization. Shanghai Library, as the collection organization for *Hu's Genealogy of ShangchuanMingjing*, is not only a name, but also an entity. The Shanghai Library not only has its full name "上海图书馆" but also its abbreviation "上图" in Chinese. It also contains the address information of "No.1555, Middle Huaihai Road, Shanghai". Most importantly, it is given a universally unique identifier "http: //data.library.sh.cn/entity/organization/11v6pvzycw 5419sy". So whenever inputting "上海图书馆" or "上图" or their traditional Chinese characters, the entity itself will be located via this unique identifier, and all the genealogy documents of this entity by the collection organizations can be found. Therefore, regardless of the attribute values of this entity inputted, in traditional Chinese or simplified Chinese, the research results keep the same. Above all, this description information exists enduringly at the underlying data, independent of the function and logic of system and can keep their semantics during the process of cross-platform and cross-system transmission and exchange.

The query is realized by using SPARQL retrieval language. SPARQL is the dedicated query language of RDF data and directly oriented to the search for knowledge. SPARQL is irrelevant to physical storage structure, but only related to the internal knowledge logic of data itself (Harris & Seaborne,2013).

The family's living place name is the same case. It no longer exists in the form of a character string, but instead as an entity which corresponds to a real existing place. This entity not only

has different names, but also contains GIS information such as longitude and latitude, and can be precisely pinpointed on the map. In view of this, we can develop the discovery and exploration functions based on a map. In addition, the more accurate browsing is fulfilled via another two ways. One is called "timeline map" browsing, and the other is called "map circle" browsing. For "timeline map" browsing, users can drag the timeline and discover how the created genealogy during a period of time is distributed on the map. For "map circle" browsing, by drawing a scope on the map, all the genealogies of the place within this region will be presented, and details are shown in Figure 5.

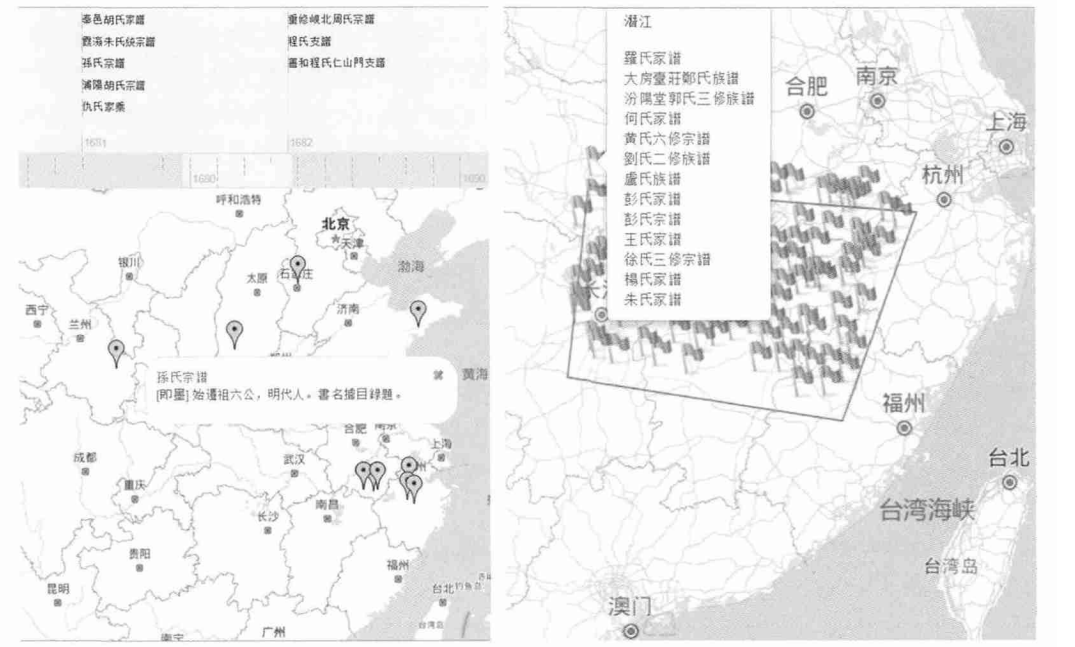

Figure 5. Genealogy knowledge discovery based on temporal-spatial correlation.

2.4.3 Platform optimization

The platform is open to users for mutual communication and interaction. It supports researchers, students, civil societies and other users from related domains to leave messages and feedbacks, modify data, and contribute knowledge. These activities will add the value and realize the proliferation of data during the use.

(1) Feedback and communication

For readers who hold the Shanghai Library card or the registered user through online registration system of the Shanghai Library can directly log onto the platform. Users can raise questions on certain genealogies by writing feedbacks and comments. Questions include inquiries on the bibliographic information of this genealogy, such as title or creator, or opinions and comments on

the genealogy contents such as information about the ancestor or migration condition. Different opinions on the data such as genealogy title, creator, and the family's living place name can be first raised and amended on the platform after the investigation, verification and acceptance by the experts.

(2) Data modification

Authenticated experts can modify the data after logging in the platform. After being verified, reviewed and approved by other experts, the modified data can be released, and the system will record each modification. Data that can be modified include description information of entities such as surname, the family's living place name, ancestor or celebrity as well as bibliographic information of the genealogy documents. The modification interface is automatically generated according to the definition of ontology. If the value range is another entity (such as author, the family's living place name or surname), one needs to choose the existing entity or establish a new identity. That is to say, if the author needs to be modified, the attribute values of this entity such as name, courtesy name, pseudonym name, dynasty, and events should be revised instead of the character string of the author's name. After modifying and saving the data, the system will record which attribute is modified, when and by whom it is modified, and the values before and after the modification. The front interface will present the new outcome after the data is released after experts' approval.

3 The publication and consumption of linked open data

3.1 Open access to public data

During the cleaning process of genealogy data, we first extract the vocabulary data from the existing metadata records, then further standardize and supplement them, and finally form the standard vocabulary list. For example, the time information used to describe the living times of the ancestor or celebrity is recorded by Chinese historical chronology. However, the information used to describe the creation time of the genealogy document is recorded by the A.D. chronology. A small part of the incomplete genealogies can only be recorded by dynastic chronology roughly. In order to unify the statements of time and realize the functions of search results ranking and timeline location, we have to etablish the correspondence between Chinese historical chronology and A.D. chronology. Therefore, we arranged the data of "Chinese historical chronology" from 841 B.C. to the Republican period of China. Considering that the correspondence between chronologies can be applied not only to genealogy, but also to other historical document resources, and serves for the Shanghai Library as well as other institutions having the similar demands, we released them to the public on our website http://data.library.sh.cn in the form of linked open data. The Restful API technology, the most commonly used in linked data consumption technology, is utilized to provide open data interface for program invocation (Xia & Liu, 2013). Developers will be able to call this

API and conduct the conversion between two chronologies by simply register an API Key on the website. The following is the API invocation method of inputting Ming dynasty and returning to the starting year of Ming Dynasty:

http：//data.library.sh.cn/time/data/明?key=yourAPIKey，返回数据为：“1368-1644”

In addition, the “Geographical Name Glossary” describes the family’s living place name and the “Organizations Directory” locates the collection organizations of the genealogy. These data in *The General Catalog of Chinese Genealogies* are taken as the paper reference index with vocabulary only. For the purpose of making better use of this geographical information, we made a further supplement. For example, for the word “Shanghai”, we added information such as “country”, “administrative region” and its “longitude & latitude”. Moreover, we added “full name” , “abbreviation” and “address” for its collection organization. Thus the function of locating it on the map can be realized and the associated relationship between data in terms of geographic latitude will be enriched. To better realize the purpose of openness and association of data on the internet, these datasets are released to the public in the form of linked open data.

3.2 Open access to genealogy data

The data in the genealogy knowledge services platform are differentiated by the entities such as surname, ancestor or celebrity, clan temple title, family’s living place name, collection organization, genealogy bibliographic data (such as title, creator, compiling time, edition information, and collection information) . Among them, place name and collection organization are extracted from metadata and released as public datasets after standardization and supplement. These data depend on the genealogy knowledge services platform and are based on *The General Catalog of Chinese Genealogies.* They take in the knowledge from experts in this field and update constantly. These updated and recognized data by experts will manifest themselves in the open data in real time. Moreover, these data can be shared and used by the whole society and new service can be created without invading personal privacy and violating laws and regulations.

The open access to genealogy data is through the website “http://data.library.sh.cn” and is based on technologies such as Http URI content negotiation, Restful API, Sqarql Endpoint. Developers can attain the genealogy data on the platform by using Http URI content negotiation and Restful API. Arbitrary combination of title, author, surname, ancestor’s name, the family’s living place name, clan temple title, collection organization name, creation time, key words in abstract are taken as input parameters, and all the matching RDF data of documents will be returned and RDF data related to these entities through the linked URI of surname, person, place, time, and orgnization also can be accessed to with the content negotiation function. RDF data will be outputted in the standard form of JSON-LD recommended by W3C and this will facilitate the programming for developers. The following is the Restful invocation method of accessing all the Xia’s genealogy

data whose place is Macheng city:

http：//data.library.sh.cn/jp/data/familyName=夏&place=麻城?key=yourAPIKey

For developers who are proficient in using SPARQL query language, they can use Sqarql Endpoint to conduct more complicated query and data access.

The website "data.library.sh.cn", as the open data platform by the Shanghai Library, will continue to release different kinds of terminology, standard archive, and collection bibliographic data on the internet and provide various data consumption interface for developers to call so as to facilitate the bibliographic control, authority control and data co-construction and sharing based on the internet.

4 Effect, problem and prospect

The genealogy knowledge services platform makes use of the modern information technologies, and presents experts' findings and the knowledge in our brain on the platform to share with the public. Resources and various statistical analysis as well as visualized research tools are used to make the resources play their full value. The Shanghai Library took the lead among domestic libraries to launch the linked data technologies for open data access. The platform has attracted broad attention once online. *China Culture Daily, Wen Hui Bao, and Xinmin Evening News* reported the news. Hundreds of genealogy research enthusiasts and various civil research societies participated in the platform evaluation and gave their feedbacks.

We came across some difficulties during the process of constructing genealogy knowledge services platform. The biggest problem is the mapping of and the conversion from the existing metadata records to the genealogical ontology framework. The existing metadata were recorded according to the traditional literature-oriented indexing method. But the genealogical ontology emphasizes both the literature properties and the content properties. So we need to scatter the original metadata records and map them to the genealogical ontology framework. Although the records of *The General Catalog of Chinese Genealogies*is are quite authoritive, some of the data are still inconsistent. As a result, it brings some difficulties and obstacles for data conversion. For example, the inconsistency in the abbreviation of collection organizations is one of the problems. The Chinese University of Hong Kong Library is among the collection information of genealogy documents in *The General Catalog of Chinese Genealogies.* The name is sometimes abbreviated as "香港中大" or "香港中文大学" in Chinese. The kind of problem is not exceptional, and it will influence the recall ratio to a large extent. Our solution is to revise all the "香港中文大学" into "香港中大" according to the organization index of the book. By using this method, no matter which word they input "香港中大" or "香港中文大学图书馆", users will find out all the genealogy resources collected by the Chinese University of Hong Kong Library. However, this solution is merely an afterward remedy. Another resolution is to understand the inconsistent problem in existing data in advance (This is obviously a very demanding task). We take "香港中大" and "香

港中文大学” as the abbreviations for “香港中文大学图书馆” (Chinese University of Hong Kong Library). During the process of data conversion or data query, the genealogy collections of “香港中大” and “香港中文大学” will be linked to the entity of “香港中文大学图书馆”.

Another problem which is even tougher is the inconsistent records between the former place names and the present ones. For illustration, “Suzhou” sometimes is recorded by its ancient name of “Wu County” in literature. However, in the current map, we can only locate the place’s current name of “Suzhou”, while “Wu County” is unable to locate. To solve the problem, “Wu County” should be taken as another property for the entity “Suzhou”. Since this kind of situation is ubiquitous, we need to depend on the “Geographical Name Glossary” which supports the correspondence between the past and the present. To construct and maintain the glossary, we need to depend on a more professional institution. However, it is also an arduous task for libraries to clarify the variations of one place name during different periods in history.

Open data has become routine rather than exceptional internationally, and it has received increasing attention in China. Genealogy, regarded as one of the most mature and influential resources by the Shanghai Library, is the leading collection implemented for open data practice. In the future, more resources will be released to the public on the “Shanghai Library Historical Documents and Data Services Platform”. One of our motivations to promote the open data of historical document resources such as genealogy is to invite more libraries and related institutions to join us to avoid redundant construction and wasting of resources, and to enhance the library value. Recently, Shanghai Library will launch a competition on data development and application based on open data interface. We truly believe that the value of data lies in its open access. Users are the real masters of the platform. Only by fostering the full awareness of open data and relying closely on users, can the library innovation remain dynamic.

References

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly*,32(4):399−418.

Cyganiak, R.,Wood, D., & Lanthaler, M. (2014).RDF 1.1 concepts and abstract syntax. Retrieved April 12, 2016, from https：//www.w3.org/TR/2014/REC-rdf11-concepts-20140225.

Fan, L. J. (2011). Influence and utilization of crowdsourcing in libraries(众包对图书馆的影响及其运用).*Library Development*(图书馆建设),(1):89−92.

Harris, S., & Seaborne, A. (2013). SPARQL 1.1 query language. Retrieved April 12, 2016, from https：//www.w3.org/TR/sparql11-query.

Library of Congress. (2015). BIBFRAME 2.0 items. Retrieved April 12, 2016, from http：//www.loc.gov/bibframe/docs/pdf/bf2-draftspecitems-10-29-2015.pdf.

Liu, W.，Zhang, C. J.，& Xia, C. J. (2015). Authority control for the Web(万维网时代的规范控制).*Journal of Library Science in China*(中国图书馆学报),(3):22−33.

Mitchell, E. T. (2013). Library linked data: Research and adoption. *Library Technology Reports*:10−15.

Sauermann, L., & Cyganiak, R.(2008). Cool URIs for the semantic Web. Retrieved April 12, 2016, from

https：//www.w3.org/TR/cooluris.
Wu, J. Z. (2015). Knowledge is fluid: New challenges to the publishing and library circles(知识是流动的：出版界与图书馆界的新课题).*Library Journal*(图书馆杂志)，34(3):4−11.
Xia, C. J., & Jin, J. Q. (2015). On the application of W3C's RDB2RDF standards(从关系数据库到关联数据：W3C标准应用探析).*Library Journal*(图书馆杂志),(5):85−94.
Xia, C. J., & Liu, W. (2013). Technologies and implementation of consuming linked data(关联数据的消费技术及实现). *Journal of Academic Libraries*(大学图书馆学报),(3):29−37.

—*Translated by* author *from* 中国图书馆学报, 2016, no.3
Revised by ZOU Yongli