Research Article **Open Access**

Xia Cuijuan*, Liu Wei, Zhang Lei

Implementation of a Linked Data-Based Genealogy Knowledge Service Platform for Digital Humanities

https://doi.org/10.2478/dim-2018-0005
received December 24, 2017; accepted March 13, 2018.

Abstract: Linked data is becoming a mature technology as a lightweight realization of the Semantic Web, as well as a way of facilitating knowledge reorganization and discovery. As a use case and start point, based on linked data technology, a genealogy knowledge service platform was implemented by the Shanghai Library for providing knowledge discovery and open data services. This article explains the design and development of the Genealogy Knowledge Service Platform, describes the method and process of the implementation, and introduces four examples of how the platform helps users to discover questions, raise questions, and solve questions for their research, to explain how Linked Data can be used in Digital Humanities.

Keywords: Digital Humanities; Linked Data; Genealogy.

1 Introduction

Digital Humanities (DH) is a buzzword that can be difficult to define. DH is "a multidisciplinary field, undertaking research at the intersection of digital technologies and humanities" (Warwick, C., 2016). DH aims to establish applications and models. It is not only a new type of research method that uses information technologies as tools to build new applications and models for studying the humanities but also contributes to the development of computer science. Meanwhile, it also studies how information technologies exert influence on cultural heritages, human memory structures, libraries, archives, and digital culture.

***Corresponding author: Xia Cuijuan,** Shanghai Library, Shanghai, China, E-mail: cjxia@libnet.sh.cn
LiuWei, Zhang Lei: Shanghai Library, Shanghai, China

The typical applications of DH are, for instance, textual analysis of classic literature or historical documents, such as William Shakespeare's works and *Dream of the Red Chamber* (one of China's four great classic novels); studies on the evolution of social conventions in combination with geographic information system (GIS); and studies on language formation and development among newborns through investigation of large amounts of recorded or real-time video data. Remarkable achievements have been made in terms of the scale, depth, and breadth of research topics.

In recent years, a number of DH research institutions have been established, such as the Alliance of Digital Humanities Organizations and the Society for Digital Humanities. Many universities have set up DH research centers, such as the Stanford Humanities Laboratory and the Department of Digital Humanities of Kong's College London.

With the continuous development of digital libraries over the past 2 decades, a large amount of monographs, periodicals, newspapers, archives, ancient books, genealogies, photographs, music, audios, and videos have been digitized. A vast mass of standard and highly structured metadata records have been generated at the same time. The research methods used by humanists have been fundamentally changed by relying on the construction of digital libraries. Digital libraries are constructed to scan, catalog, and organize library collection resources. The emphasis is on the digitized forms of resources, and the aims are to support document management with computers while providing reference service for readers at the same time. When the construction reaches a certain scale, this kind of document-focused management and service encounters bottlenecks. The search results reach dozens of pages and thousands of records, and yet the readers must browse the full text page by page. As a result, it becomes quite difficult for them to discover the relevant facts or knowledge from such vast resources. Now, various modern digital technologies, such as "Cloud Computing", "Big Data", and the "Semantic Web", provide methods that can analyze the content and the knowledge within based on logical relationships among the digital resources on

a large scale. And they also can provide accurate service oriented toward mining hidden knowledge. It could save manpower and time when organizing and promoting the use of resources. Moreover, it could even generate new knowledge in the utilization of resources.

Data and technologies are two major pillars of DH. The metadata and digitized full-text documents left over from the construction of digital libraries lay the foundation for data on DH services. Linked Data is a mature technology as a lightweight realization of the Semantic Web. The implementation of DH on the basis of Linked Data has become a trend and has been widely used in many projects abroad.

2 Literature Review of DH Research and Linked Data Applications

Many new DH projects, such as the digitization of ancient books and literature databases, have been emerging in Europe and America. The projects "Mapping the Republic of Letters" by Stanford Humanities Research Center (Coleman, et al., 2016) and "The Stanford Geospatial Network Model of the Roman World" are the classic cases. Libraries in China have begun to realize that the construction of digital libraries has reached a significant scale. New opportunities and breakthroughs need to be sought from DH. The theme of the 2014 Information Technology for Library (IT4L) seminar was "Digital Humanities and Semantic Technologies". Professors, librarians, and practitioners from Shanghai Library and Fudan University, as well as from libraries in the USA and New Zealand, discussed the concepts and technologies of DH and exchanged ideas on achieving DH through semantic technologies.

As one of the technological measures of DH, Linked Data technology has already attracted extensive attention and application. In recent years, almost every DH conference has had at least one session concerned with the topic of "Digital Humanities, RDF, and Linked Data". The annual international Digital Humanities 2015 conference was held in collaboration with the third international Linked Open Data in Libraries, Archives and Museums (LODLAM) Summit (Page, 2016). Projects that applied Linked Data technologies to realize DH services emerged one after another, such as "Digitized Manuscripts to Europeana (DM2E)" (Baierer, 2016) and "Linked Humanities project" (Sztyler, et al., 2016), which are comparatively full-fledged ones.

The DM2E project is presided over by the School of Library and Information Science, Humboldt-Universität zu Berlin in Germany and supported by research institutions such as the University of Mannheim and the Open Knowledge Foundation. The project has two primary goals: 1. The transformation of various metadata and content formats into the Europeana Data Model (EDM) and loading of these data into Europeana, the European Digital Library. 2. The reorganization and publishing of a Linked Data data set based on Linked Data technologies so as to support DH. An additional goal is the development of tools for data display, processing, and integration, as well as for utilization on other projects. DM2E has published a Linked Data data set and opened it under CC0 license by default. However, data providers are also allowed to select from other specific open access licenses. The data set contains a variety of historical resources, including manuscripts and related humanities resources such as archive items, letters, books, and journal articles. The DM2E model is designed based on the EDM, which is itself very generic in order to represent the various kinds of historical resources in Europeana provided by museums, libraries, archives, and galleries all over Europe. The DM2E model, as an application profile of EDM, passes on the extendibility and inclusiveness of EDM. The data set follows the four principles of Linked Data and conforms to the five-star scale for Open Data. The data set is available for academic search through a RESTful API sufficient for data-consuming applications that can be programmatically obtained and integrated into other DH service platforms. This supporting platform enables scholars to access the data set contents, create annotations, and refer to other resources. In addition, the platform supports version control mechanisms and retains scholars' operational history. The link discovery tool SILK is utilized to add external links to the data set automatically. These functions are realized mainly by two customized tools. The first one is a faceted browser, "OmNom", which allows scholars to navigate and exploit DM2E collections along several dimensions. The other tool is a semantic annotation tool called "Pundit", which allows users to enrich Web pages with semantically structured data. Annotations in Pundit are organized based on the "Open Annotation Data Model" and encoded into a machine-readable format RDF and stored in a triple store, where they are consumable via Simple Protocol and RDF (Resource Description Framework) Query Language (SPARQL) (or a dedicated REST API) and available for program debugging.

"Linking and Populating the Digital Humanities", known more simply as the "Linked Humanities Project", was a 2-year-long project (from 2012 to 2014), which was funded by the National Endowment for the Humanities (NEH) and the German Research Foundation (DFG) and presided over by Indiana University Bloomington. The primary

motivation for the project was to build tools for integrating and maintaining DH data and constructing associations between these data so as to serve humanities studies via the enhancement of sharing and reuse of data. This project developed a framework called Linked Open Data Enhance (LODE), which is applicable to other DH projects. LODE features an explorer component, a linking component, and an enhancement component. Concrete use cases such as the "Indiana Philosophy Ontology (InPhO)" and the "Stanford Encyclopedia of Philosophy (SEP)" have been explored. Furthermore, Web services have been developed to integrate with external data repositories to promote the application of a philosophy ontology. The LODE project has also been applied to Jewish culture and history. By building links between academic reference resources and publishing linked data, data-consuming services are made available on the Web. Meanwhile, under the semantic annotations framework built by the DM2E project, annotations can be added to these resources, which enable humanities researchers to add more links and enrich the contents.

In terms of genealogy resources, libraries – especially domestic libraries – are inclined to consider genealogy as a kind of document resource. Therefore, MAchine-Readable Cataloging (MARC) and *Dublin Core (*DC) metadata are used to index document features and a small number of content features. A search is conducted via keyword matching to fields such as genealogical title, creator, edition, ancestor, clan temple title, or residence. Typical cases are the Oriental Library of the National Diet Library (NDL) in Japan, "Zhonghua Xungen Wang (http://ouroots.nlc.gov.cn)" created by the National Library of China, and genealogy repositories collected by various domestic institutions, which, similar to the original genealogy database of Shanghai Library, do not support discovery of content and knowledge. This precious resource for humanities research has not been fully excavated by today's genealogists. DH methods and technologies, especially the developed Linked Data technology, however, will help improve the current situation.

In recent years, ontology-based semantic technology has been utilized to refactor genealogy data formatting abroad. Successful examples are "FamilySearch.org" by the Genealogical Society of Utah (GSU) and "ancestry.com", which not only provide a keyword search function that describes document features but also enable users to explore a genealogy resource and its member relationships by featuring contents such as temporal–spatial correlation and kinship. One innovative case is the "Kindred Britain" network, a Stanford Libraries-based Digital Humanities project. Kindred Britain presents a beautifully visualized network of nearly 30,000 famous people throughout 1,500 years of British history. The network, on which connections between different individuals (including family relationships of blood or correlations generated from the same time and space) are vividly shown, can help scholars quickly discover new knowledge from massive amounts of data.

3 The Application of Linked Data in Genealogy Knowledge Service Platform

3.1 The Implementation of Four Linked Data Principles

When it comes to data, the DH requires knowledge in fine-grained units, organized semantically and demonstrated visually. The first and second principles of Linked Data require that the entire "Thing" needs to be identified and located via hypertext transfer protocol (HTTP) uniform resource identifier (URI) (Sauermann., &Cyganiak 2008), laying a foundation for Web-based authority control. The third principle of Linked Data calls for the use of Resource Description Framework (RDF) (Cyganiak., Wood &Lanthaler 2014) as the abstract data model, revealing the relationships between different resources as much as possible. RDF is a scientific knowledge organization method, which takes triples (statements consisting of "subject-verb-object") as the basic data units. This differs from the traditional practice of matching one document to a corresponding metadata record. One piece of a metadata record can thereby be split into multiple triples. A triple is the description for a certain piece of knowledge, data, or fact. It possesses independent description logic and can make the fine-grained knowledge unit a reality. The fourth principle of Linked Data places emphasis on the inherent correlation between data. Machine-readable correlation of data can also make the quest for semantic knowledge organization a reality. Accordingly, genealogy data reorganized and published on the basis of Linked Data can fulfill the data requirements of the DH.

3.1.1 (1) Use HTTP URIs as Names For all the Things

Here are the diverse entities of HTTP URIs in Shanghai Library's genealogy data. They are taken as the unique identifiers and universal locators of resources. By using these entities, the foundation has been laid for achieving the objective of Web-based authority control.

Genealogy Document: http://data.library.sh.cn/jp/resource/work/8y9p7s2euppwnerq

Person: http://data.library.sh.cn/jp/entity/person/qxj915pkohm96unn

Family Name: http://data.library.sh.cn/authority/familyname/68n959cf8zdfkz3v

Place: http://data.library.sh.cn/entity/place/tk5s4pej6linq9tr

Temporal: http://data.library.sh.cn/authority/temporal/4alljneqiivh5691

Organization: http://data.library.sh.cn/entity/organization/brvqlrg8y55v1b5q

3.1.2 (2) Knowledge Organization Based on Ontology and RDF

RDF data is a triple composed of "subject-verb-object". Its subject is a resource object, and the object can be either a string value or another resource object. The verb expresses the relationship between subject and object and is derived from a strictly defined ontology. The triple also embodies independent semantic knowledge units, which can be recognized by a machine after encoding into RDF serialization formats, such as RDF/XML, RDF/Turtle, JSON-LD. and so on. Following is the description of the Hongwu Period of the Ming Dynasty in RDF/Turtle format:

```
<http://data.library.sh.cn/authority/temporal/3rwxdjxxfz5bhff9>
a shl:Temporal;

        bf:label "明洪武";

        shl:monarch "太祖";

        shl:monarchName "朱元璋";

        shl:reignTitle "洪武";

        shl:dynasty <http://data.library.sh.cn/authority/temporal/
yex4deivsad41p9q>;

        time:intervalDuring

<http://data.library.sh.cn/authority/temporal/yex4deivsad41p9q>

        shl:beginYear 1368;

        shl:endYear 1398.
```

3.1.3 (3) When Accessing the Resource URI, Useful Information will be Provided in an RDF Serialization Format

Genealogy Knowledge Platform must support a content negotiation mechanism. This means that when a human being accesses a resource URI, the platform will return an HTML page, but when a program accesses that same resource URI, the RDF data will be returned in the JSON-LD format.

If a program accesses a resource URL and uses. json in its request to ask for the return data in the JSON format (e.g., http://data.library.sh.cn/jp/entity/person/qxj915pkohm96unn.json), then the following result is returned:

{„result“:{„data“:{„@id“:“http://data.library.sh.cn/jp/entity/person/qxj915pkohm96unn“,“@type“:“http://www.library.sh.cn/ontology/Person“,“label“:[{„@language“:“cht“,“@value“:“星華”},{“@language”:”chs”,”@value”:”星华”}],”relatedWork”:”http://data.library.sh.cn/jp/resource/work/8y9p7s2euppwnerq”,”roleOfFamily”:”http://data.library.sh.cn/jp/vocab/ancestor/xian-zu”,”familyName”:”http://data.library.sh.cn/authority/familyname/rn3hurvwucnb24pb”,”@context”:{“familyName”:{“@id”:”http://xmlns.com/foaf/0.1/familyName”,”@type”:”@id”},”relatedWork”:{“@id”:”http://www.library.sh.cn/ontology/relatedWork”,”@type”:”@id”},”label”:”http://bibframe.org/vocab/label”,”roleOfFamily”:{“@id”:”http://www.library.sh.cn/ontology/roleOfFamily”,”@type”:”@id”}}}}}

3.1.4 (4) Provide as Many URIs Linking to Other Resources as Possible to Support Information Discovery

In an RDF description of genealogy documents, HTTP URIs are used to refer to entities such as family name, person (creator, ancestor, or celebrity), place (residence), time (dynasty in the Chinese historical calendar), and organization (contributor/owner of the collection). The properties defined in an ontology are used to express the relationships between these entities and the documents. When exploring the description data, these related entities' URIs link to more detailed information about the person, place, time, and event. For example, when visiting the dynasty URI below, the year span of the dynasty from 1368 AD to 1644 AD will appear.

http://data.library.sh.cn/authority/temporal/yex4deivsad41p9q

```
<http://data.library.sh.cn/jp/resource/work/oyz2f36kouez9jdy>
a bf:Work;

shl:place <http://data.library.sh.cn/entity/place/
t3tec8y1oy2j3kjc>;//地

shl:temporal <http://data.library.sh.cn/authority/temporal/
yex4deivsad41p9q>;//时

bf:creator < http://data.library.sh.cn/jp/entity/person/
etrd44w3m3g1vncn>;//人

bf:subject <http://data.library.sh.cn/jp/authority/titleofancestraltemple/6biawgp5dbdm9hkm>;

bf:subject <http://data.library.sh.cn/authority/
familyname/4feibkhldtiroeu3”//姓氏

<http://data.library.sh.cn/jp/resource/instance/
apz8aio37wj6y524> a bf:Instance;

bf:category <http://data.library.sh.cn/vocab/binding/xian-zhuang>;

bf:edition <http://data.library.sh.cn/vocab/edition/mu-zi-huo-ben>;

bf:extent “五册“;

shl:temporalValue 1936, “1936年”;

bf:instanceOf <http://data.library.sh.cn/jp/resource/work/
oyz2f36kouez9jdy>;

<http://data.library.sh.cn/jp/resource/item/fgxpi3bc8km3r672>
a bf:Item;

bf:heldBy<http://data.library.sh.cn/entity/organization/
uoqz22aqnemd3idn>;//机构

bf: itemOf<http://data.library.sh.cn/jp/resource/instance/
apz8aio37wj6y524>;

shl:description “漳州市臺工辦林嘉書（複印本，存卷一、三、五）”;
```

3.2 The Technical Framework of Genealogy Knowledge Service Platform

Linked data technology is implemented in building the genealogy knowledge services platform because it can do the following: 1) organizes knowledge based on a domain-based conceptual system (ontology), instead of documents; and 2) describes and retrieves knowledge via the RDF abstract data model, which can be expressed as RDF triples using grammatical constituents (subject, verb, object). Using existing data proofing and

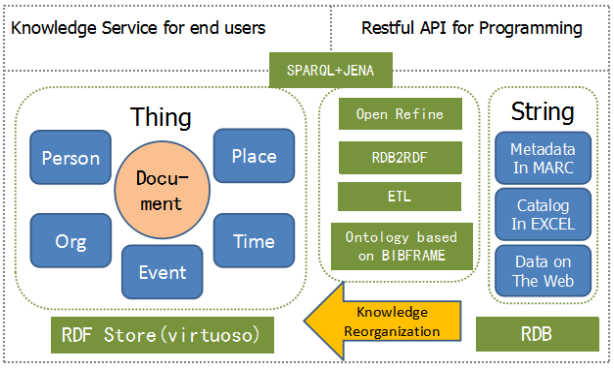

Figure 1: The Technical Framework of the Platform

knowledge mining tools to support the maintenance and updating of knowledge, users are allowed to access data within the document instead of the whole document. On the other hand, with the widespread and profound application of linked data in libraries, a set of technologies, methodologies, and processes for metadata, ontology, and RDF data conversion, RDF data storage and querying, and data visualization has matured, which can meet the requirements of bibliographic control and authority control, data reuse and sharing, and knowledge organization and discovery in Web environment.

There is a technical framework that supports knowledge reorganization and provides knowledge service – not just document retrieval and reading, thus fulfilling the requirements of DH (Figure 1). The framework is based on other existing frameworks, models, standards, and tools and was built using the following process.

First, we gathered data from metadata records in the SQL Server of Shanghai Library’s genealogy collection (http://search.library.sh.cn/jiapu/), data in EXCEL format from the Chinese Genealogy Catalog covering 597 organizations, and data on the Web (such as stories about family names, geographic data from geonames.org, and relationships among people from Google Knowledge Graph) to supplement the library data. The supplementary data from these various sources all needed to be integrated together and transformed into RDF.

Second, in order to make “Things not Strings” from the family names, places, and people in the genealogy documents, we needed to design a data model (ontology) that will provide each Thing with its own HTTP URI and structural RDF data. The genealogy ontology we created uses the three-tier bibliographic model of LOC’S BIBFRAME 2.0 (http://www.loc.gov/bibframe/docs) as the core data model. It reuses some terms from existing ontologies and vocabularies, such as schema.org, GeoNames, the World Wide Web Consortium (W3C)’s Time Ontology, and so

on. In total, it contains >40 classes and 110 properties. The ontology model and the associated vocabularies are published on the Web in RDF/XML serialization format (http://gen.library.sh.cn:8080/ontology/).

Next, we needed to dissect, clear, and transform the data we had gathered according to the ontology. The data in both the SQL Server and the EXCEL Catalog can only be accessed in the relational database with the structure of "Record-Field-Value". Thus, the data needs to be transformed from RDB to RDF, one process of which is called RDB2RDF. There are two kinds of open source conversion tools that can be applied to accomplish the RDB2RDF process: one is "DB2Triples", which supports the W3C's RDB2RDF standards (Xia & Jin, 2015); the other is "OpenRefine", which is a very popular open source tool for data profiling and cleaning. Both tools support the creation of a mapping between the tables and fields in RDB on the one hand and the classes and properties in the genealogical ontology, definition of the rules of HTTP URI generation, and the automatic generation of RDF data on the other. We also developed an ETL (Extract, Transform, and Load) tool, which can gather data from different data sources and then clear data and transform it into RDF data.

Finally, we needed a way to store the RDF data. As the first library to provide Linked Open Data services in China, our team worked out a solution to implement Linked Data, with the RDF data stored in a triple store instead of a traditional relational database. The triple store can well support the RDF query language SPARQL and the Semantic Web development framework Jena.

For user services to genealogy researchers, a variety of data visualization tools or plugins (such as SIMILE Timemap, BaiduEcharts, and AutoNavi map) can be used. For open data services to application programmers, three kinds of data-consuming technologies are available: HTTP URI content negotiation, RESTful API, and SPARQL Endpoint.

3.3 Characteristics of the Genealogy Knowledge Services Platform

3.3.1 (1) Quantity and Range of Data

The data on our platform comprises the largest collection of Chinese genealogy documents in existence in the LAM (Library, Archive, Museum) community; we cleaned the catalog data of 60,000 genealogy documents held by 597 organizations all over the world, including libraries, archives, and museums. Then, we transformed this catalog data into RDF data according to the classes and properties defined by our custom-designed genealogy ontology based on BIBFRAME (Library of Congress, 2015). The "Work-Instance-Item" model of BIBFRAME can support bibliography control very well, allowing our platform to be used for a global cataloging of genealogy documents. Users can get all of the information about one genealogy document which is held by different organizations. This helps users enormously in cases such as when they cannot find the document in the Shanghai Library but know where to find it in other another organization's collection.

3.3.2 (2) Authority Control of Data

To maintain good data quality, we use the method of Authority Control used in libraries. The genealogy ontology mentioned herein contains four classes (FamilyName, Person, Place, and Temporal), each of which serves as a Controlled Name for a specific kind of data. Figure 2 shows the entire abstract model of the Shanghai Library's data infrastructure for DH, which is also suitable for genealogy (Xia,2016).

Chinese family names can behave very differently from their Western counterparts. It is common for many people to share the same last name, and some family names are changed from their original form from ancient times. This means that there are some very important properties that are necessary to identity a family name, including relationships among family names. Moreover, the pronunciation of an everyday Chinese word changes when it is used as a family name. We have a family name knowledge base with 628 family names to organize this information using the "shl:FamilyName" class and its properties. 'shl' is the prefix of Shanghai Library's

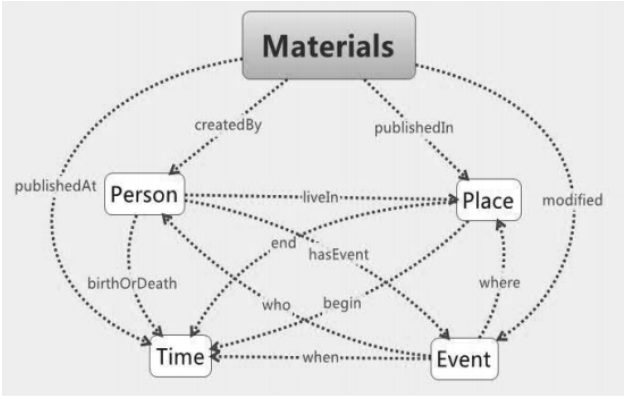

Figure 2: Shanghai Library's data infrastructure for Digital Humanities

namespace. Classes and Properties with this prefix are defined by our own.

For people names, the same person may have multiple names and different people may have the same name. We have a controlled people names knowledge database (http://names.library.sh.cn) with ~800,000 people – so far – based on the "shl:Person" class and its properties. Every name is controlled with HTTP URI and RDF data to describe it.

For place names, the process is also very complicated. The name of a place may have changed many times throughout history. Because of the fact-based description principle, many place names are recorded using their ancient names in metadata records of genealogy. Similar to the people names, there are cases where different places have the same name. Authority control is needed to resolve these problems by merging the different names of the same place together and differentiating the different places with the same name whenever readers request genealogy records for a certain place. Based on the geonames.org ontology, we developed a knowledge database with >1,800 place names. Every name has an HTTP URI and RDF data set containing geographic data and its relationship among place names.

For a time, in ancient Chinese books including genealogy documents, time was expressed using a special Chinese calendar year with the King's name in various dynasties. Based on the W3C's Time Ontology, we designed a temporal ontology and developed a knowledge database to map the Chinese calendar years to the Western calendar years from the Tang dynasty to the Qing dynasty.

These knowledge databases of Chinese place names and time periods are accessible on the Web (http://data.library.sh.cn).

3.3.3 (3) The Data Services for researchers

There are several playable user interfaces on the genealogy knowledge service platform. First, an interface with a time controller and interactive map helps users find genealogy documents from a certain period and geographical region (Figure 3). Some users may not know the exact place name where their ancestors lived, so they cannot find genealogy documents according to ancestors' residence, which is a very important access point for data retrieval. To address this difficulty, we developed a map tool that allows users to draw a polygon on the map outlining a general area, thus enabling one to search genealogy documents without knowing the exact place name. This tool is very useful and popular according to the feedback of end users (Figure 4).

The most important information represented by a genealogy document is the personal information of family members over several decades. However, it is difficult to read it all at once in a static document. So, we developed an interactive family tree for showing a pedigree table as a Web page. Users can explore details about the birth and death, important life events, the individual's parents and children, and other relationships among people. Figure 5 shows several generations from among the total 42 generations of Hu Shi's Family from the Tang Dynasty to the early 20th Century. Users can explore the whole family

Figure 3: Time and map interface.

Figure 4: A data retrieval tool by drawing a polygon on the map.

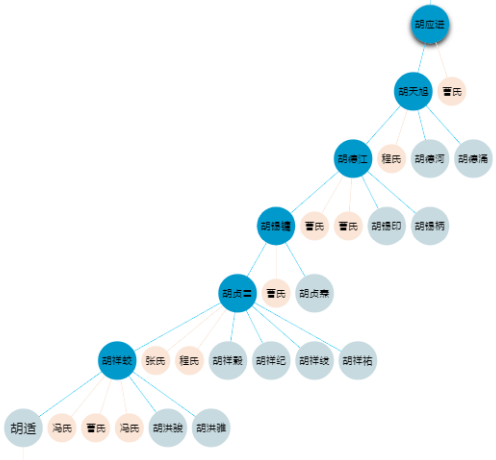

Figure 5: An interactive family tree.

tree by double-clicking the nodes and read the detailed information of a person in the tree, such as birthday and death day, important events and stories, and the relationships with other people.

4 Use Cases for DH Services

4.1 Data in Small Granularity for Visualization to Tell Stories and Help Users Ask Questions

Family migration is one very important type of data recorded in genealogy documents. Studying the migration route of a family, a region, or a certain time period has been an essential subject in many branches of the humanities, such as demography, anthropology, and history. A visual example of one family's migration is presented in the map "Shangchuan Ming Jing Hu's migration map", which traces the movements of the ancestors of Hu Shi, a Chinese cultural celebrity (Figure 6). Among the Shanghai Library's genealogy collection, there are six genealogies that are related to Hu's family of Ming Jing. The metadata of two genealogies records that Hu Shi is one the most famous people of his family. It also records that the first ancestor of Hu's family is "Hu Changyi".

According to the genealogy ontology model, one of the properties possessed by first ancestor is migration. For each migration, there are migration properties of person, time, and place, i.e., who, when, from where, and to where.

Property values (entities – not strings), such as person, place, and time, are extracted from the migration event recorded in genealogy metadata according to the data structure defined by the event ontology, which is one part of the whole genealogy ontology. Described by the RDF data format, these entities are connected into the migration event one after another and can link the ancestors who have the same ancestor in their migration events. The following are RDF data fragments of a migration event:

```
//the migration event
<http://gen.library.sh.cn/Migration/1> a shlgen:Migration;
        bf:eventAgent <http://gen.library.sh.cn/Person/16607>;
        shl:temporal <http://gen.library.sh.cn/Temporal/14>;
        shl:locality <http://gen.library.sh.cn/Place/1129>;
        shl:originalLocality <http://gen.library.sh.cn/Place/203>.
//the person of the migration event
<http://gen.library.sh.cn/Person/16607> a shlgen:Person;
        foaf:name "胡昌翼";
        shlgen:courtesyName "宏远";
        shl:pseudonym "眉轩";
        shl:temporal <http://gen.library.sh.cn/Temporal/14>;
        shl:description "唐昭宗幼子".
//the original locality of the migration event
<http://gen.library.sh.cn/Place/203> a shlgen:Place;
        bf:label "长安";
        shl:city "西安市";
        shl:province "陕西省";
geo:long " 108.93";
geo:lat " 34.27".
//the target locality of the migration event
<http://gen.library.sh.cn/Place/1129> a shlgen:Place;
        bf:label "考川";
        shl:county "婺源县";
        shl:city "上饶市";
        shl:province "江西省";
        shl:town "紫阳镇";
        shl:village "考水村";
        geo:long "117.757787";
        geo:lat "29.27642".
//the time when the event happened
<http://gen.library.sh.cn/Temporal/14> a shlgen:Temporal;
        shl:dynasty "五代后唐";
        shl:beginYear "923"^^xsd:int;
        shl:endYear "936"^^xsd:int.
```

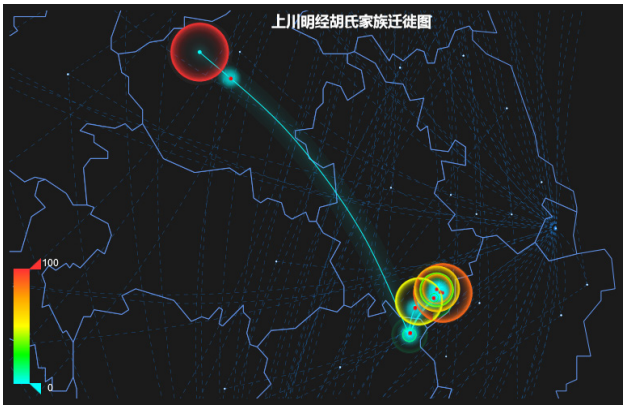

Figure 6: Migration map of Shangchuan Ming Jing Hu's family

The following figure visually presents the migration route from the Five Dynasties period and the later Tang Dynasty to the late Qing Dynasty and the early Republic of China via use of the E-charts visualization tool from Baidu to automatically generate a migration map from the above RDF data. The map vividly reveals the family's story throughout history. It tells the complete story of the migration of "Hu Changyi", the first ancestor of Shangchuan Ming Jing Hu's family, from the capital city Chang'an to Kaochuan village in Wuyuan during the late Tang Dynasty, as well as the migration routes of Hu's offspring from Kaochuan village to diaspora like Jixi County in Anhui Province.

Thus, the system automatically uses migration events as links between ancestors from different families, who, through the study of multiple genealogies, have been found to share the same first ancestor. The migration route and the offspring's destinations are automatically generated by the system according to migration time and place.

On the one hand, readers will therefore directly acquire knowledge from different documents and organize it in accordance with some predefined subject so as to answer a question that the user already knew they wanted to ask. On the other hand, the system will also help users to discover and raise questions. The above migration map visually unveils to us that Hu Changyi, the first ancestor of Shangchuan Ming Jing Hu's family, experienced a remote migration journey from Chang'an to Kaochuan. During a time when transportation was very inconvenient, readers may wonder why Hu undertook such a migration and what was the motivation behind it? However, guided by the links from the above map, readers will discover descriptive metadata in the related genealogy documents. In the abstract introducing one such documents, readers will find answers to their questions: "The first ancestor Changyi was the youngest son of Emperor Zhaozong of Tang and Empress Ho in Tang Dynasty. He was born in 904 AD. Changyi moved to Kaochuan village in Wuyuan to escape calamity and then changed his family name into Hu which was the surname of his adoptive father Hu Sangong." This reveals the fact that Hu Shi's first ancestor, Hu Changyi, was the son of the last emperor of the Tang Dynasty. In other words, Hu Shi is the descendant of Tang emperors. This kind of dramatic story will inspire users to probe into genealogy resources.

4.2 Authority Control at the Web-Scale to Increase Precision and Recall for Domain Specific Research

On the Internet, the goal of authority control is to realize concept-based description and matching instead of character-based matching. Linked Data technologies based on HTTP URIs, the RDF data model, and ontologies constitute such an ideal solution. When a place is identified uniquely by an HTTP URI, given an ontology-defined concept, and described by RDF data, it no longer exists solely in the form of a character string but also as an entity corresponding to a real existing place. That entity, which serves as an online surrogate for the real-world thing, thereby contains abundant RDF description data and can not only be located on the Internet, but also be read and processed by machine – all owing to the RDF description that includes the current place name, ancient place name, alias, province or city to which the place belongs, as well as the longitude and latitude. As a result, the problems of different names for the same place and different places with the same name can be resolved, and the entire hierarchy of names describing a place, both sub- and superordinate, can also be analyzed.

The first use case presented the migration route of a family. However, the system can also support the analysis of migration within a certain geographical region or during a specific period of time. In the description metadata, there is an entity called "place" (also "family's living place"), which refers to the family's residence at the time the genealogy was compiled. The entity is generally described at the county level. Take "Sichuan Province" as an example. When all the subordinate counties of Sichuan Province are made relevant to the "Sichuan Province" search through the property "gn:parentADM2", the system will quickly conduct dynamic clustering for all the related genealogies whose family's living places were located in Sichuan Province. Thus, we can get a result of 956 instances of genealogy documents. Furthermore,

the migration to Sichuan Province can be clustered, which reveals that there are 221 instances of genealogies recording the migration from "Macheng City, Hubei Province" to "Sichuan Province". This number is close to a quarter of the total genealogy documents, a proportion that can be considered compelling evidence for the topic, "Hu-Guang Fills Sichuan", the largest immigration in China's history. In the book, *Emigration History of China* by Professor Ge Jianxiong, the famous historical geography researcher mentions that plenty of records of migration from "Xiaogan Village, Machen" to "Sichuan Province" during the Ming and Qing periods have been discovered in genealogies rather than in the official history and chorography (systematic description and mapping of regions or districts). Therefore, it is not an exaggeration to claim that genealogy is an indispensable resource, along with official history and chorography.

```
<http://data.library.sh.cn/entity/place/topq2nlfiugtg964>

a shl:Place;

bf:label "四川"@cht"四川"@chs;

shl:abbreviateName "川";

    shl:country "中國"@cht"中国"@chs;

owl:sameAs <http://www.geonames.org/1815285>;

gn:lat       30.67;

gn:long 104.07.

<http://data.library.sh.cn/entity/place/1v1smmtsmqg29rlf>

    a shl:Place;

bf:label "樂山"@cht"乐山"@chs;

shl:province "四川省";

gn:parentADM2 <http://data.library.sh.cn/entity/place/
topq2nlfiugtg964>;

... ...
```

4.3 Statistics and Analysis of Data for Evidence-based Research

The pedigree table is the paramount part of a genealogy. It records – in detail – information (such as birth and death, marriage, migration, and lifetime events) on all the male members of a family, as well as their spouses and children, from the first ancestor to the period when the genealogy was compiled. Among this information, there are abundant data, facts, and knowledge, but all in the form of scanned images at present. The Shanghai Library has attempted to transform these pedigree tables from images into structured data. The following RDF description of "Hu Shi" describes the basic information of the sort described above, apart from relationships with other resources, such as relationships between people (e.g., father and son, husband and wife), as well as between people and literature. These relationships are expressed using the properties defined by the ontology as the predicates and then encoded as triples in machine-readable Turtle format.

```
<http://data.library.sh.cn/jp/entity/person/uqwn21pa7iah9p87>
a shl:Person;

    shl:name "洪骅";

    foaf:name "胡适";

    shl:courtesyName "适之";

    foaf:familyName <http://data.library.sh.cn/authority/
familyname/rvmgzfsec8os93mv>;

    shl:orderOfSeniority "3";

    shl:birthday "光绪辛卯十一月十七未时";

    shl:description "考派美国留学生名适字适之事迹见学了生光绪
辛卯十一月十七未时娶江氏生光绪庚寅十一月初八辰时";

shl:relatedWork <http://data.library.sh.cn/jp/resource/work/
jklhb5c3ga1rvxe3>;

shl:roleOfFamily<http://data.library.sh.cn/jp/vocab/ancestor/
xian-zu>;

rel:spouseOf <http://data.library.sh.cn/jp/entity/person/
kvop3qdif1okxhoa>;

    rel:childOf <http://data.library.sh.cn/jp/entity/person/
mc7khkqgql3rpceb>.
```

The structured data can also be applied to statistics and analysis. For instance, from the pedigree table of *Shangchuan Ming Jing Hu's Genealogy,* the names of 8,915 male family members and 4,733 of their spouses are separated out. Using statistical analysis, it is found that although the surname "Li" is one of the most common surnames in China, it does not exist among the 4,733 spouses. Instead, the surname "Cao" is the most common one, for there are 982 spouses with this family name (Figure 7). The reason is mainly explained by a very important principle: marriage between people of the same surname

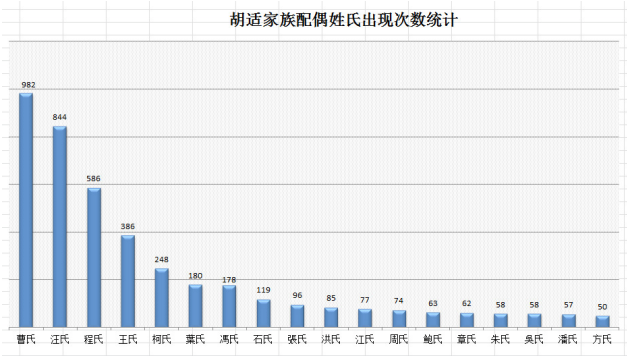

Figure 7: Statistics on the frequency of spouses' family names in the pedigree table of *Shangchuan Ming Jing Hu's Genealogy*.

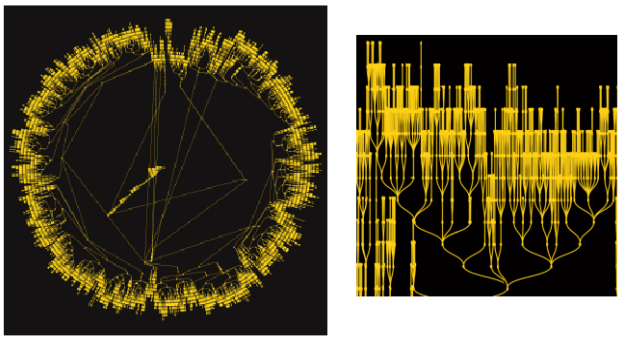

Figure 8: Artistic Design of Ming Jing Hu's Family tree

was prevented at that time. These data provide strong evidence for the statement in the aforementioned abstract (Section 4.1) indicating that Hu Changyi is the offspring of Li Ye, known as Emperor Zhaozong of Tang. Obviously, the statistical evidence is more vivid and persuasive than plain words.

4.4 Opening Data to the Public for More Creative Use of Library Data in Humanities

Linked data is designed for Open Data. The consumption of genealogy data and authority data within the Genealogy Knowledge Service Platform is through our website (http://data.library.sh.cn) via HTTP URI content negotiation, RESTtful API, or SPARQL Endpoint. Developers can access genealogy bibliographic data such as record title, author, organization name, creation time, family name, ancestor's name, the family's living place name, clan temple title, and key words from each record's abstract. All of these data can be used as input parameters, and all the matching RDF data of documents will be returned in JSON-LD, the format recommended by the W3C. JSON-LD facilitates programmatic access to data for developers. For example, the following is the RESTtful invocation method for accessing all the genealogy data of the "Xia" family name associated with the place name "Macheng City": http://data.library.sh.cn/jp/data/familyName=夏&place=麻城?key=yourAPIKey

For developers who prefer to use SPARQL query language (Harris, &Seaborne 2013), they can access the SPARQL Endpoint (http://data.library.sh.cn:8890/sparql) to run more complex queries against the data.

As the open data platform of Shanghai Library, data.library.sh.cn will continue to publish different kinds of data (e.g., vocabularies, controlled names, data sets, and tools) on the Web and provide various data consumption interfaces for developers to call on, so as to facilitate Web-based bibliographic control, authority control, as well as data co-construction and sharing.

In 2016, we held an open data application development contest based on these Linked Open Data consumption technologies. One hundred and thirty-nine developers joined the contest and designed apps that showcased many great ideas on how to use genealogy data and to mash up that data on the Web. By exposing open data on the Web, we found a way to link end users, domain experts, third-party developers, and librarians.

So far, a number of developers and designers from other branches of the humanities have been glad to use our open genealogy data to fulfill their own research requirements. For example, Xiang Fan and Zhu Shunshan from the Academy of Arts and Design, Qinghua University, have used our pedigree table data to design an artistic exhibition of the Ming Jing Hu family tree (Figure 8). It is very different from Figure 5. In this case, >10,000 people show up in a single figure. The relationships among those people are designed as an artistic work, which can tell stories such as what time period the most family members lived in? What branch of the family has the most members?

5 Conclusion

Genealogy Knowledge Service Platform, a project of the Shanghai Library, is an attempt and exploration toward realizing the goals of DH research using Linked Data technologies. By making use of existing metadata, Shanghai Library aims to mine facts, knowledge, and data to develop knowledge graphs based on the relationships among entities and to help users explore what they need. Libraries, as providers for DH services, need not only to answer their users' questions but also need to inspire them to raise new ones. In addition, it is libraries' responsibility

to offer more convenient methods and tools that will help users explore deeply, yet efficiently, a large amount of historical data.

The cases expounded on in this study are based on just one family's migration map and pedigree table. If *all* genealogy documents become structured and semantic data, the change in how we are able to utilize these resources will be revolutionary – a panoramic, visual demonstration and statistical analysis of the relationships among people on the basis of massive data, combining both time and space. However, structuring and defining migration information and pedigree tables is a large-scale work, especially the large number of ancient places and village names that need to be correlated with historical geographic names. The Shanghai Library is now developing data set tools, such as the "Geographical Name Glossary", to deal with the task of disambiguating place names. As for processing the data within pedigree tables, the platform will develop functions such as "Crowdsourcing" so as to gain more support and involvement from the whole society.

Since the Genealogy Knowledge Service Platform went online, there have been as many as 6,000 clicks a day, and the project has been reported on by dozens of newspapers. People became fans of us, and they themselves built a user community with hundreds of people from many places both within and outside of China. They discuss interesting stories about genealogy together and also give us wonderful ideas on how to correct the data and ways to improve our platform. Our next steps will be to update our platform to allow users and researchers to generate data more conveniently, upload their own genealogy documents, and share their research results.

Furthermore, because of some obsolete ideas reflected in genealogies (such as "tuft hunting", "airing the good and punishing the evil", and "preference for sons"), the reliability of existing conclusions based on genealogy data needs to be further verified. It should be clarified, therefore, that DH should be considered a set of technologies and methods rather than the ultimate goal, a supplement rather than a substitute for humanities research. Technologies furnish platforms and tools, but it is the users who make the final decisions. In other words, technologies can assist users in discovering problems, but they must search out the evidence for themselves to resolve these problems and form conclusions of their own. Equally important is that DH contributes to improving user experience and encourage patrons to make use of library resources. Finally, the realization of DH based on Linked Data will help more third-party data services to discover library resources and will increase both the volume and quality of their utilization.

References

Baierer, K. (2016).*DM2E: A Linked Data Source of Digitised Manuscripts for the Digital Humanities*. Retrieved from http://www.semantic-web-journal.net/system/files/swj831.pdf.

Coleman, N. (2016). *Digging into the Enlightenment: Mapping the Republic of Letters* Retrieved from http://www.clir.org/pubs/reports/pub151/case-studies/enlightenment.

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

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

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

Page, K.. (2016). *A Humanities Web of Data: Publishing, Linking and Querying on the Semantic Web*. Retrieved from http://digital.humanities.ox.ac.uk/dhoxss/2014/HumData.html.

Sauermann L., & Cyganiak R. (2008). *Cool URIs for the semantic Web*. Retrieved from https://www.w3.org/TR/cooluris.

Sztyler, T., J. Huber, J. Noessner, J. Murdock, C. Allen, and M. Niepert (2016). LODE: *Linking Digital Humannities Content to the Web of Data*. Retrieved from http://arxiv.org/pdf/1406.0216v1.pdf.

Warwick, C., Terras, M. (2012). Digital humanities in practice. In Nyhan, J. (Eds.). London: Facet Publishing.

Xia, C. (2016). Building a Digital Humanities Platform by Using Linked Open Data Services. *Journal of Library and Information Science*, 43(1), 47 -70.

Xia, C.J., & Jin J.Q. (2015). On the application of W3C's RDB2RDF standards. *Library Journal*, (5), 85-94.