

*The Implementation of
Historical and Humanities
Big Data Platform of
Shanghai Library*

Xia Cuijuan



DC2020



历史人文大数据平台
Digital Humanities Platform of Shanghai Library

目录

Contents

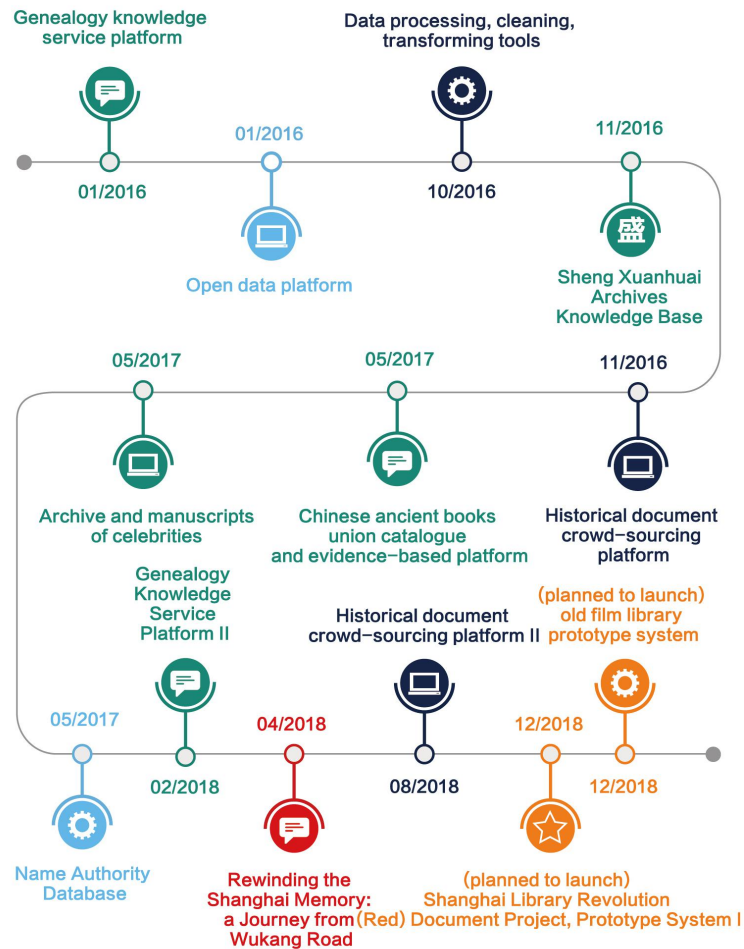
- 1 壹 • *Background and problems*
- 2 贰 • *Solutions*
- 3 叁 • *Technologies and Methods*
- 4 肆 • *Planning*

1 壹

Background

DH Projects of Shanghai Library and the difficulties during the development process

DH Projects of Shanghai Library



☁ Process (from DL to DH)

- Transform all metadata records of different resources in different formats into RDF. Take authors, contributors, publishers, etc. as entities.
- Give every entity cool URI as global identifier and locator then link them together after NER and entity disambiguation.
- Enrich more semantic data for the Entities by extracting structured data from the content of resources or Open datasets on the web.
- Provide open data and authority control in a web-scale, Digital Humanities services for researchers and APIs for the third-party developers.



Difficulties

- Data Cleaning and NER. Have to correct the inconsistency of metadata records in digital library systems and deal with the entity disambiguation manually.
- Provide proper Services for researchers. Not easy to understand the specific requirements of researchers from different areas. The services of so many knowledgebases and applications need to be integrated and well designed for user need and user experience.

2 贰

Solutions

Build an infrastructure to Bridge the gaps among the process of digitalization, data production, data cleaning, NER, and LOD publishing.

Build one platform to integrate the services of multiple DH systems and meet the specific requirements of researchers from different areas.

Goals

- One portal to access all
- One engine to search all
- One KOS to integrate all
- One data hub to Mash-up all
- One infrastructure to fulfill all

The screenshot displays the '历史人文大数据平台' (Digital Humanities Platform of Shanghai Library) interface. The search term '李鸿章' (Li Hongzhang) is entered in the search bar. The results are categorized into several sections:

- 以“李鸿章”检索 人名规范库 得到 39 个结果**: Includes a portrait of Li Hongzhang and a table of his basic information.
- 以“李鸿章”检索 中国家谱知识服务平台 得到 15 个结果**: Lists family genealogy records.
- 以“李鸿章”检索 中文古籍联合目录及循证平台 得到 154 个结果**: Lists classical Chinese literature records.
- 以“李鸿章”检索 盛宣怀档案知识库 得到 5953 个结果**: Lists records from the Sheng Xuanhuai archive.
- 以“李鸿章”检索 近代报刊资源库 得到 0 个结果**: Lists records from the modern newspaper resource library.

On the right side, there is a vertical timeline of events related to Li Hongzhang, including his participation in the founding of the Shanghai Commercial Education Society and his role as the founder of the company.

<https://dhc.library.sh.cn>

Technical Architecture

- Digital objects Layer

IIIF Server, PDF Server, Image Server provide scan image management and access for all digital objects.

- LOD Layer

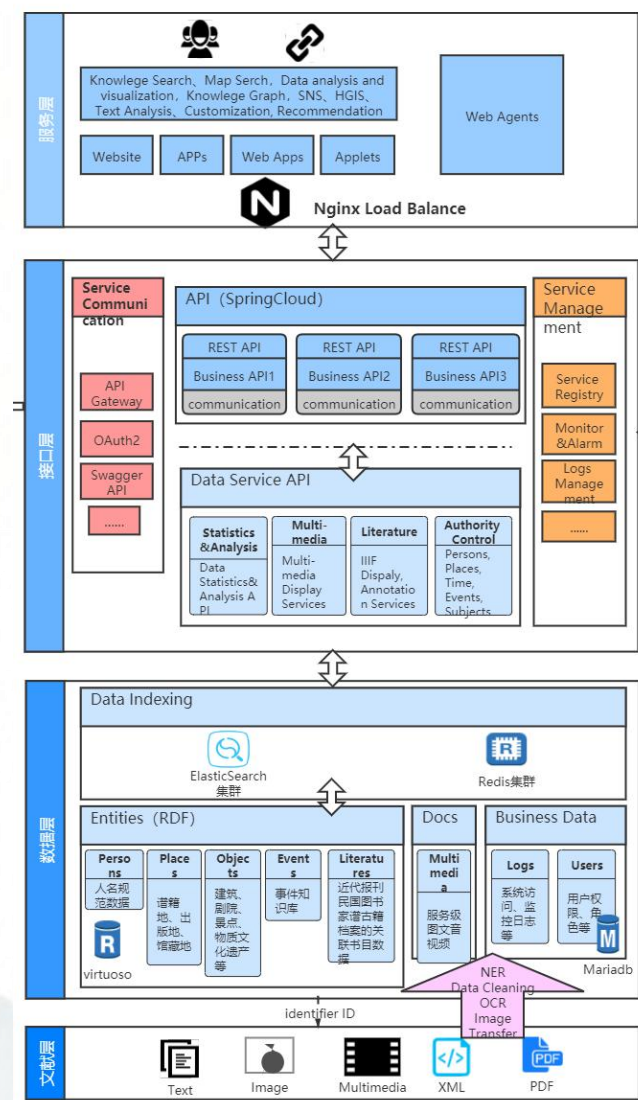
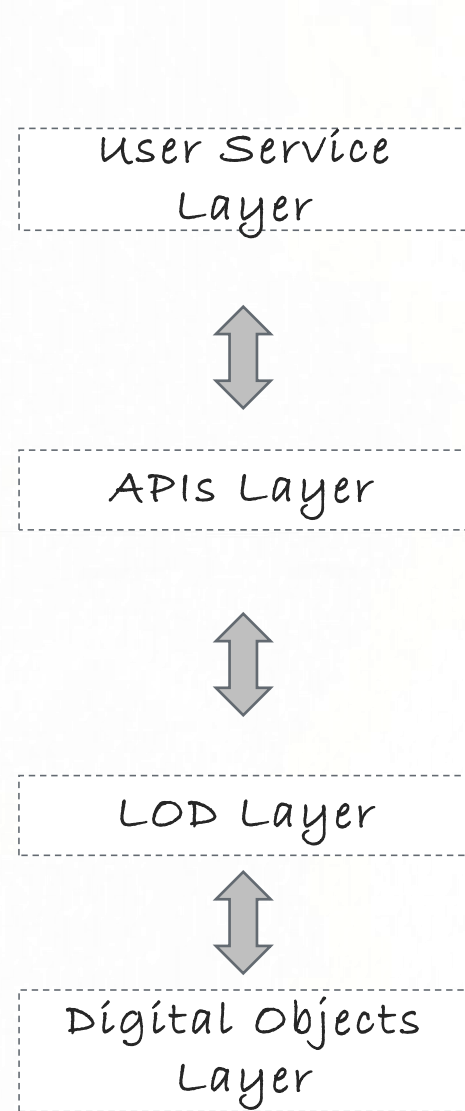
The RDF data of all Metadata and authority data is stored in RDF Store and published as LOD.

- APIs Layer

Performs data output and input between LOD layer and user service layer

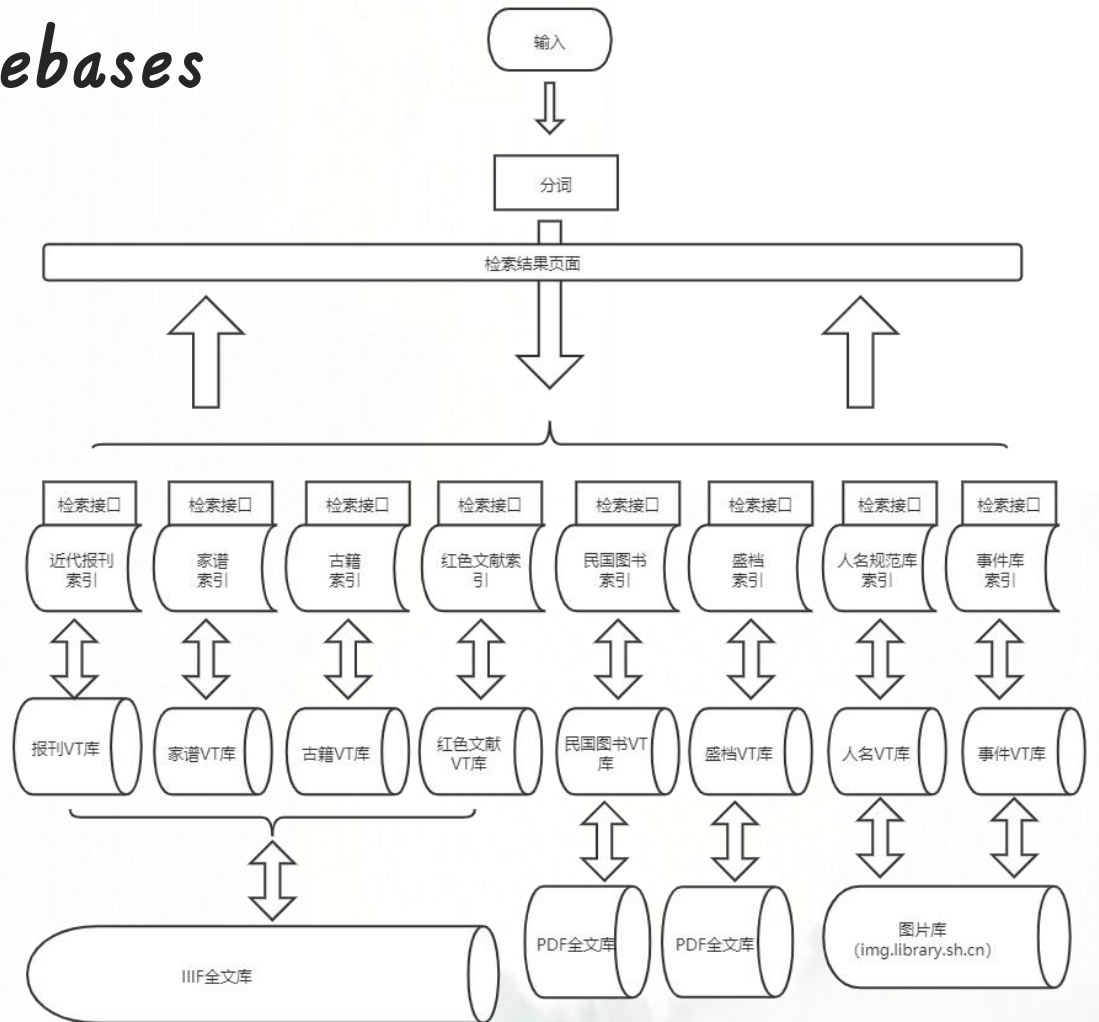
- User Service Layer

Provid navigation, search, visualization, statistics and analysis.....



Search Engine Cross Knowledgebases

- Every knowledgebase is based on LOD technologies, and the RDF data is stored in Openlink Virtuoso (VT).
- Every Digital Object (scanned image) is linked to the RDF data and displayed in the framework of IIIF based on IIP Server.
- A Federal search engine through APIs provided by Elasticsearch index of every single knowledgebase.



3 卷

Technologies

Crowd sourcing for Image2Text

LOD for metadata and authority data publishing and sharing on the web

Machine Learning for NER

IIIF for displaying and reorganization of the scanned images

Text analysis, SNS, GIS for typical DH research paradigms

Crowd Sourcing for Image2Text

- Transcription online
- Captcha



上海图书馆 SHANGHAI LIBRARY

汉冶萍公司辛亥年财产损失如何入帐请示 截止日期：2020-12-31

主题：辛亥革命；财产；汉冶萍公司

日期公元：[1912年5月26日]

日期年号：[民国元年]四月十日

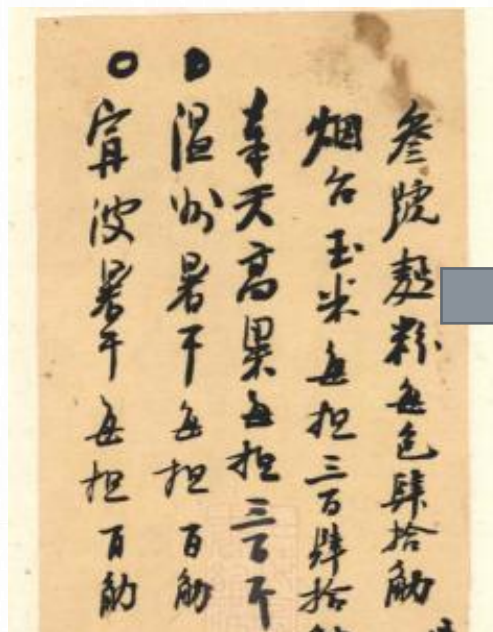
文件类型：信函

全文抄录

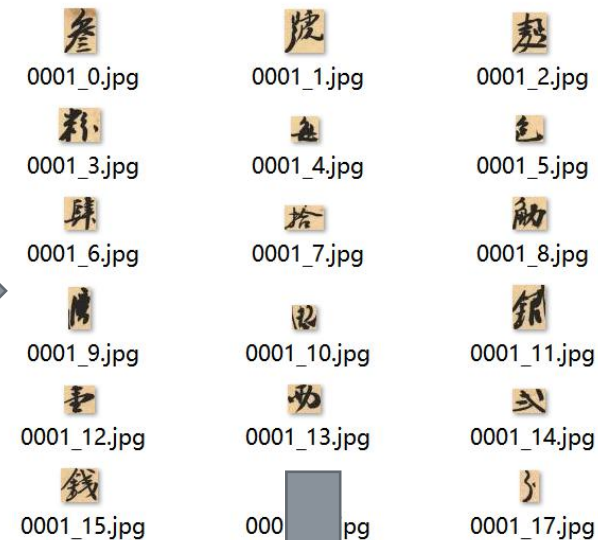
厂矿兵乱损失，辛亥年居其多数，来册仍照历界帐略造报，将来损失一项，势难一概列入。壬子帐内或由公司查明列入，或请其于册尾登明，方为正办。盖辛亥九月以后，会长远在东洋【.....】经理声明，则会长序中措词，未免无裨矣，请钧裁。四月十日

厂矿损失，未必有帐，总办须亲去看过也。只可序中带叙，俟壬子年帐内开列全叙。经理之意，尚倾向入汉口商务，请国家赔偿也。

<http://zb.library.sh.cn>



页号	块号	行号	字号	最终结果	最终结果识别次数	总识别次数	刷新次数
0001	0	0	0	叁	26	33	32
0001	0	0	1	披	6	103	1408
0001	0	0	10	傲	6	180	7167
0001	0	0	11	银	6	198	485
0001	0	0	12	壹	6	18	24
0001	0	0	13	助	6	289	173
0001	0	0	14	或	8	200	
0001	0	0	15	钱	6	8	
0001	0	0	17	3	10	15	8
0001	0	0	18	有	6	10	48
0001	0	0	19	右	6	10	20
0001	0	0	2	面	5	138	7347
0001	0	0	3	粉	10	10	36
0001	0	0	4	每	8	20	15



叁 0001_0.jpg

叁 0001_1.jpg

叁 0001_2.jpg

叁 0001_3.jpg

叁 0001_4.jpg

叁 0001_5.jpg

叁 0001_6.jpg

叁 0001_7.jpg

叁 0001_8.jpg

叁 0001_9.jpg

叁 0001_10.jpg

叁 0001_11.jpg

叁 0001_12.jpg

叁 0001_13.jpg

叁 0001_14.jpg

叁 0001_15.jpg

叁 0001_17.jpg

授权登录

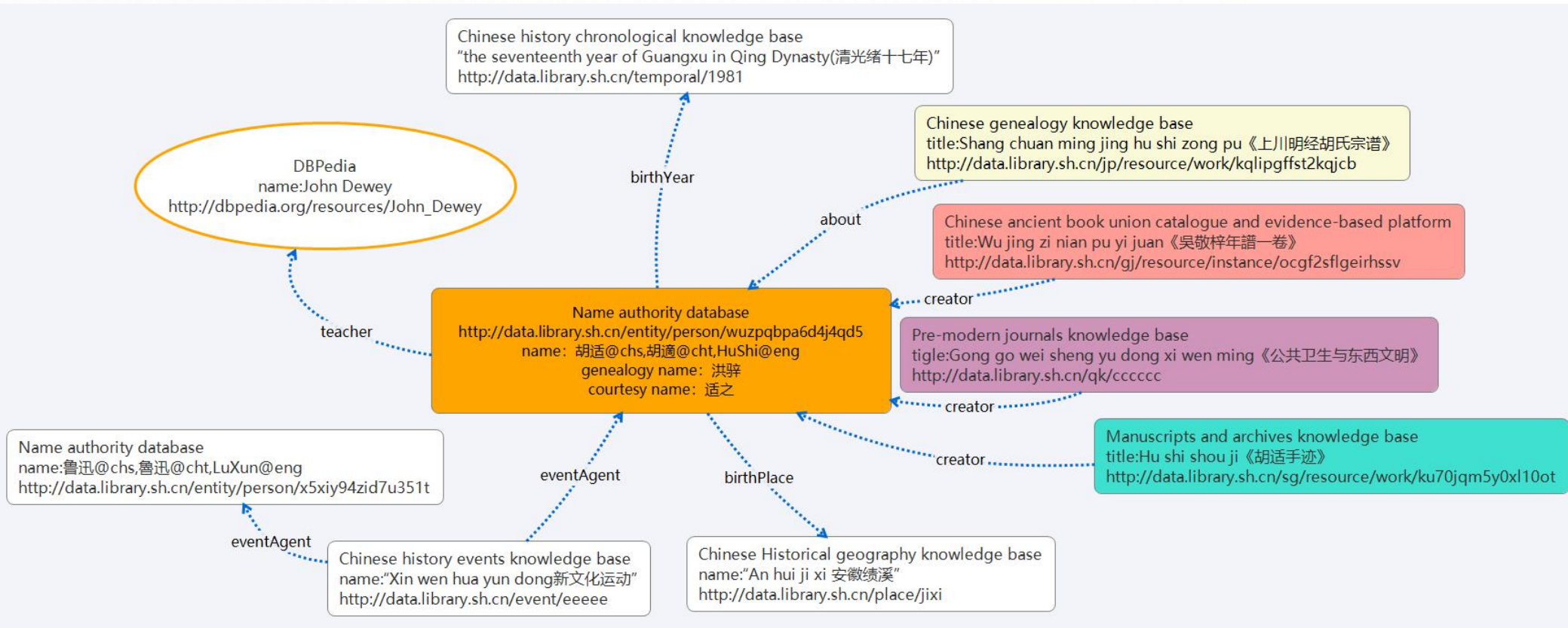
授权登录

请输入用户名/读者卡号

初始密码为办证时所使用的证件号

记住我 一周

LOD for Metadata and Authority data publishing and sharing

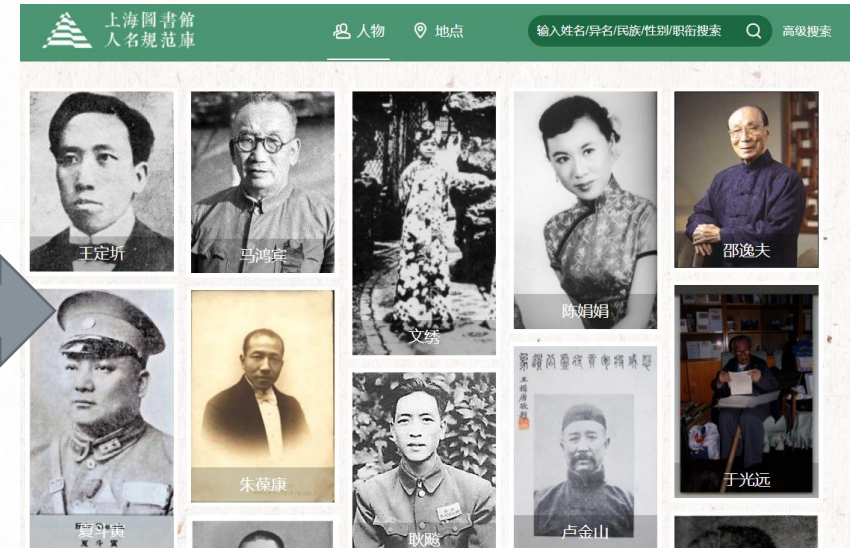


(more than 300 millions of RDF triples totally)

Machine Learning for NER

BERT (Bidirectional Encoder Representation from Transformers)

标题	主题词/关键词	事件内容	事件发生时间	事件开始时间	事件结束时间
1 千顷堂书局创办	千顷堂书局, 黄产生, 张松涛, 鲍兴华	千顷堂书局于1883年由黄产生创办, 设址于南市:	1883	1883-01-01	1955/12/3
2 光华书局创办	光华书局, 沈松泉, 张静庐, 卢芳, 《洪水》, 《光华书局于1925年由沈松泉、张静庐、卢芳创办。		1925	1925/1/1	1935/5/3
3 现代书局由洪雪帆	现代书局, 洪雪帆, 张静庐, 卢芳《现代小说》	现代书局于1927年由洪雪帆创办。编辑部 and 经理!	1927	1927/1/1	1919/8/1
4 昆仑书店创办	昆仑书店, 李达, 邓初民, 熊得山, 张正夫, 熊子民, 昆仑书店于1927年冬筹办, 1928年正式开业。由		1928	1927/10/1	1932/12/3
5 大江书铺于1928年	大江书铺, 陈望道, 施存统, 汪馥泉, 冯三昧, 《大江月》	大江书铺于1928年9月创办, 由陈望道、施存统、	1928	1928/9/1	1933/12/3
6 朝花社于1928年1	朝花社, 鲁迅, 柔石, 王方仁, 崔真吾, 许广平, 朝花社于1928年11月在上海创办, 新文学社团之		1928年11月	1919/8/6	1930/5/3
7 中国科学图书仪器	中国科学图书仪器公司, 中国科学社, 任鸿宾, 中国科学图书仪器公司于1929年6月由中国科学社		1929年6月	1929/6/1	1955/12/3
8 佛学书局于1930年	佛学书局, 王一亭, 李经纬, 范古农, 沈彬翰, 《佛学书局于1930年1月在上海创办。初时设址闸北		1930	1930/1/1	1949/12/3
9 儿童书局于1930年	儿童书局, 张一渠, 石芝坤, 潘公展, 黄仲明, 庞来	儿童书局于1930年2月由张一渠、石芝坤合资创办	1930年2月	1930/2/1	1950/12/3
10 三闲书屋于1931年	三闲书屋, 鲁迅, 《梅斐尔德木刻士敏土这图》	《三闲书屋于1931年创办, 鲁迅自办, 针对国民党	1931	1931/1/1	1937/12/3
11 《生活》周刊编辑	艾寒松, 《生活》周刊	艾寒松 (1905~1975), 原名逸尘, 又名涂尘。;1905年	1905	1905/1/1	1975
12 《立报》主笔	恽逸群, 《立报》《解放日报》《新闻学讲话》	恽逸群 (1905~1978), 原名钥勋, 字长安, 笔:1905年	1905	1905/1/1	1978
13 史量才秘书、曾任	马荫良, 《申报》	马荫良 (1905~1995), 字一民。上海松江人。1905年	1905	1905/1/1	1995
14 "副刊圣手"张慧剑	张慧剑	张慧剑 (1906~1970), 原名张嘉谷, 笔名辰子。1906年	1906	1906/1/1	1970/5/14
15 《正言报》创办人	吴绍澍, 《正言报》	吴绍澍 (1906~1976), 字雨生。上海市金山人。1906年	1906	1906/1/1	1976/6/30
16 上海《东南日报》	胡健中, 《东南日报》	胡健中 (1906~1993), 原名经亚, 字絮若, 笔名翟1906年	1906	1906/1/1	1993/9/21
17 文汇报社社长	金仲华, 《东方杂志》《中学生》《世界知识》	金仲华 (1907~1968), 幼名翰如, 笔名孟如。;1907年	1907	1907/1/1	1968/4/21
18 《立报》总编辑兼	萨空了, 《立报》	萨空了 (1907~1988), 笔名了了、艾秋颺。蒙;1907年	1907	1907/1/1	1988/10/11
19 上海《大公报》总	徐铸成, 上海《大公报》《文汇报》	徐铸成 (1907~1991) 笔名荆紫、银丝、丁宁。;1907年	1907	1907/1/1	1991/12/21
20 《新民报》副经理	邓季惺, 《新民报》	邓季惺 (1907~1995), 四川奉节人。幼习私塾 1907年	1907	1907/1/1	1995/8/21
21 上海《大公报》副	费彝民, 上海《大公报》	费彝民 (1908~1988), 1908年12月22日出生于1908年	1908	1908/12/22	1919/8/1

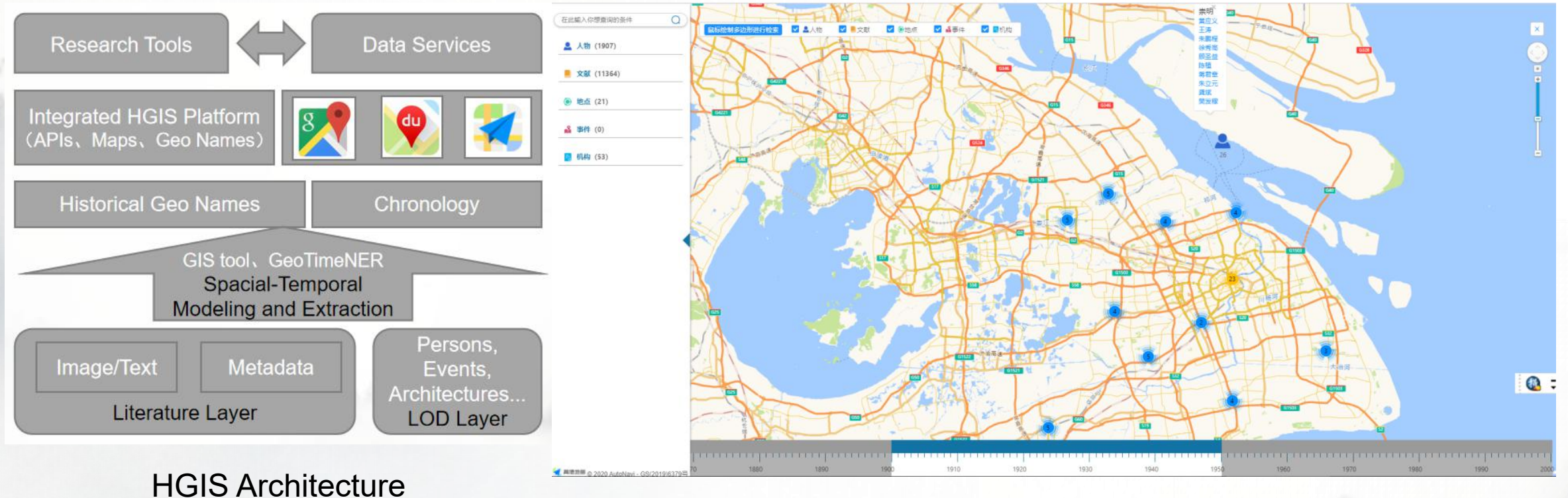


5548 person entities from 7430 records, 159 incorrect, 314 omitted, Accuracy: 91.47%

IIIF for Displaying, Sharing, and Reorganization of the Scanned Images

The screenshot displays the platform's interface for viewing a scanned newspaper page. The top navigation bar includes search and navigation icons. The left sidebar contains a list of document titles, such as "中國人為全世界人類鼻祖 (續) (p1)" and "對閱者諸君的報告(p1)". The main content area shows a newspaper page with the title "张恨水《啼笑因缘》" (Zhang Henshui's 'Tixiao Yinyuan'). The newspaper page features dense vertical columns of text and several advertisements, including one for "大廉" (Dalian) and another for "金瓶梅" (Jin Ping Mei). The bottom status bar indicates the current page is "新闻报19300316期0021版" (Xinwen Bao, March 16, 1930, Issue 0021, Page 1).

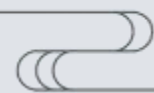
GIS for searching and visualizing big data on the map



4 肆

Planning

The future plans in the next 5 years to accomplish the Platform



Planning

2015~2025

Phase 2

- New Technologies
- New Methods
- Data and knowledge Integration
- Busyness and services integration

Phase 4

- Outreach
 - Cultural tourism
 - Digital exhibition
- Innovation Support



Phase 1

- Transform Digital library systems to knowledgebases based on LOD technologies
- DH projects development

Phase 3

- Support Specific Application scenarios
- Rebuilding and long term preservation of City memory
 - Specific fields research support

Thanks

◆ Questions and Comments ◆

Xia Cuijuan
xtykc@yeah.net



DC2020



历史人文大数据平台
Digital Humanities Platform of Shanghai Library