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

Xia Cuijuan

DC2020
Contents

1 壹・Background and problems
2 贰・Solutions
3 叁・Technologies and Methods
4 肆・Planning
Background

1

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

- 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.

Process (from DL to DH)
Linked entities and data of different knowledgebases with URIs semantically by the One Ontology Abstract Model.
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.
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

https://dhc.library.sh.cn
Historical and Humanities Big Data Platform

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

Provide navigation, search, visualization, statistics and analysis......
Historical and Humanities Big Data Platform

KOS Based on One Ontology

- Ancient Books
- Archives and Manuscripts
- Old Movies
- Genealogy
- Old Photos
Historical and Humanities Big Data Platform

🔍 **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 ElasSearch Index of every single knowledgebase.
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

http://zb.library.sh.cn
Historical and Humanities Big Data Platform

LOD for Metadata and Authority data publishing and sharing

(more than 300 millions of RDF triples totally)
Historical and Humanities Big Data Platform

Machine Learning for NER

BERT (Bidirectional Encoder Representation from Transformers)

5548 person entities from 7430 records, 159 incorrect, 314 omitted, Accuracy: 91.47%
Historical and Humanities Big Data Platform

IIIF for Displaying, Sharing, and Reorganization of the Scanned Images
Historical and Humanities Big Data Platform

Text Analysis
Historical and Humanities Big Data Platform

SNS for exploring the relationships among people
Historical and Humanities Big Data Platform

GIS for searching and visualizing big data on the map

HGIS Architecture
The future plans in the next 5 years to accomplish the Platform Planning
Historical and Humanities Big Data Platform

Planning

2015~2025

Phase 1
• Transform Digital library systems to knowledgebases based on LOD technologies
• DH projects development

Phase 2
• New Technologies
• New Methods
• Data and knowledge integration
• Busyness and services integration

Phase 3
• Support Specific Application scenarios
• Rebuilding and long term preservation of City memory
• Specific fields research support

Phase 4
• Outreach
  • Cultural tourism
  • Digital exhibition
• Innovation Support
Thanks

Questions and Comments

Xia Cuijuan
xtykc@yeah.net

DC 2020