A review of Wikidata in DH projects and some cases in the study of premodern Chinese culture

Fudie ZHAO (趙芙蝶)

DPhil Candidate, Faculty of Asian and Middle Eastern Studies, University of Oxford

fudie.zhao@sant.ox.ac.uk

Outline

1. What is Wikidata? Why is it interesting?
2. Some cases in Chinese studies
3. A mini hands-on session where we can edit one or two items on Wikidata collaboratively
4. Wikidata for a comprehensive collaborative research environment for the study of premodern Chinese culture?
5. A review of Wikidata in DH projects

1. What is Wikidata?

Wikidata and its interface

- https://www.wikidata.org/wiki/Q4604

Wikidata

- a free and open **knowledge base** that can be read and edited by both humans and machines.
- **a central storage** for the **structured data** of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others.
- support many other sites and services beyond just Wikimedia projects.
- Its content is available under a **free license**, exported using **standard formats**, and can be interlinked to other open datasets on **the linked data web**.

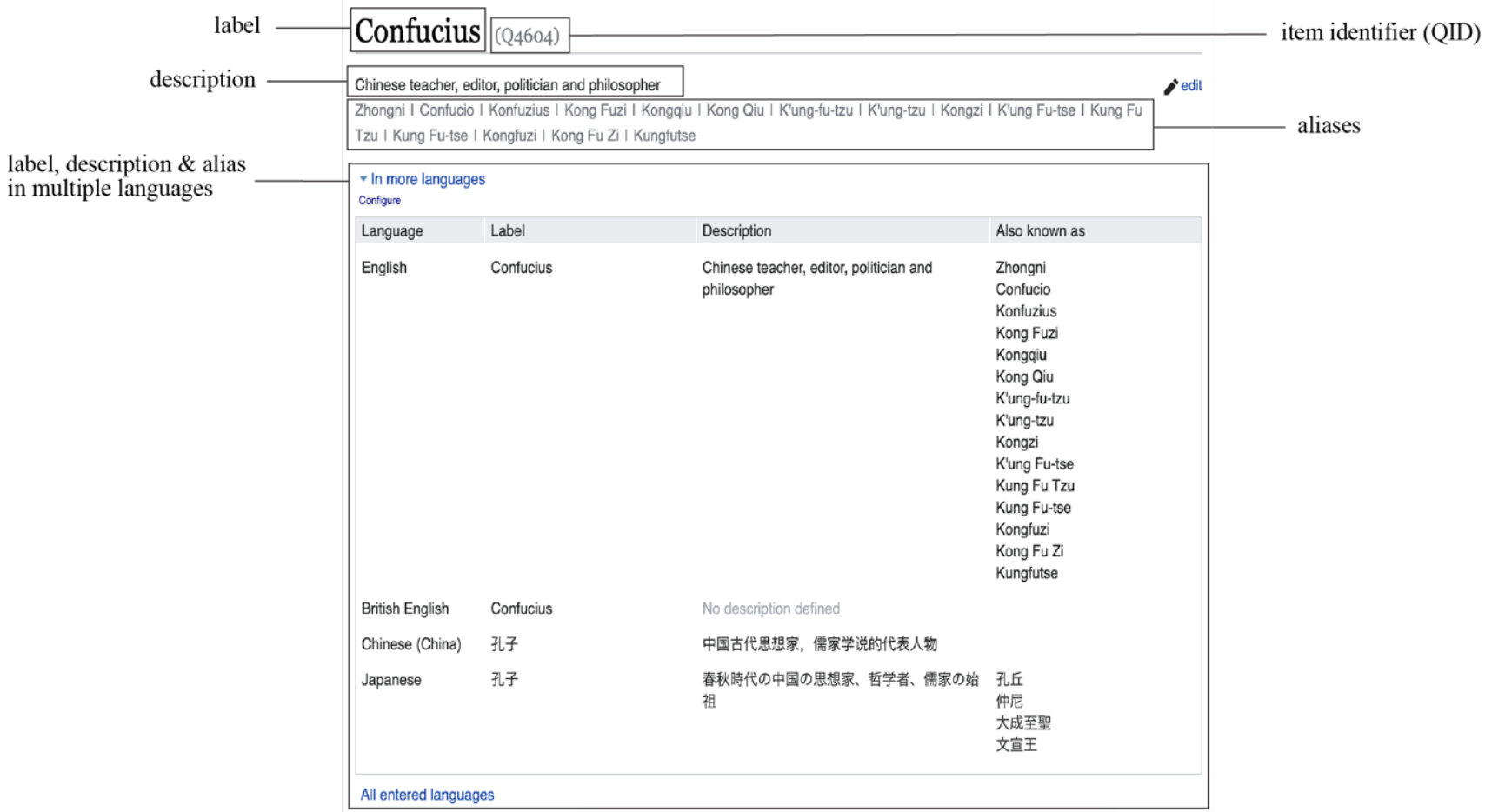
Wikidata

- A simple, user-friendly, collaboration-oriented interface

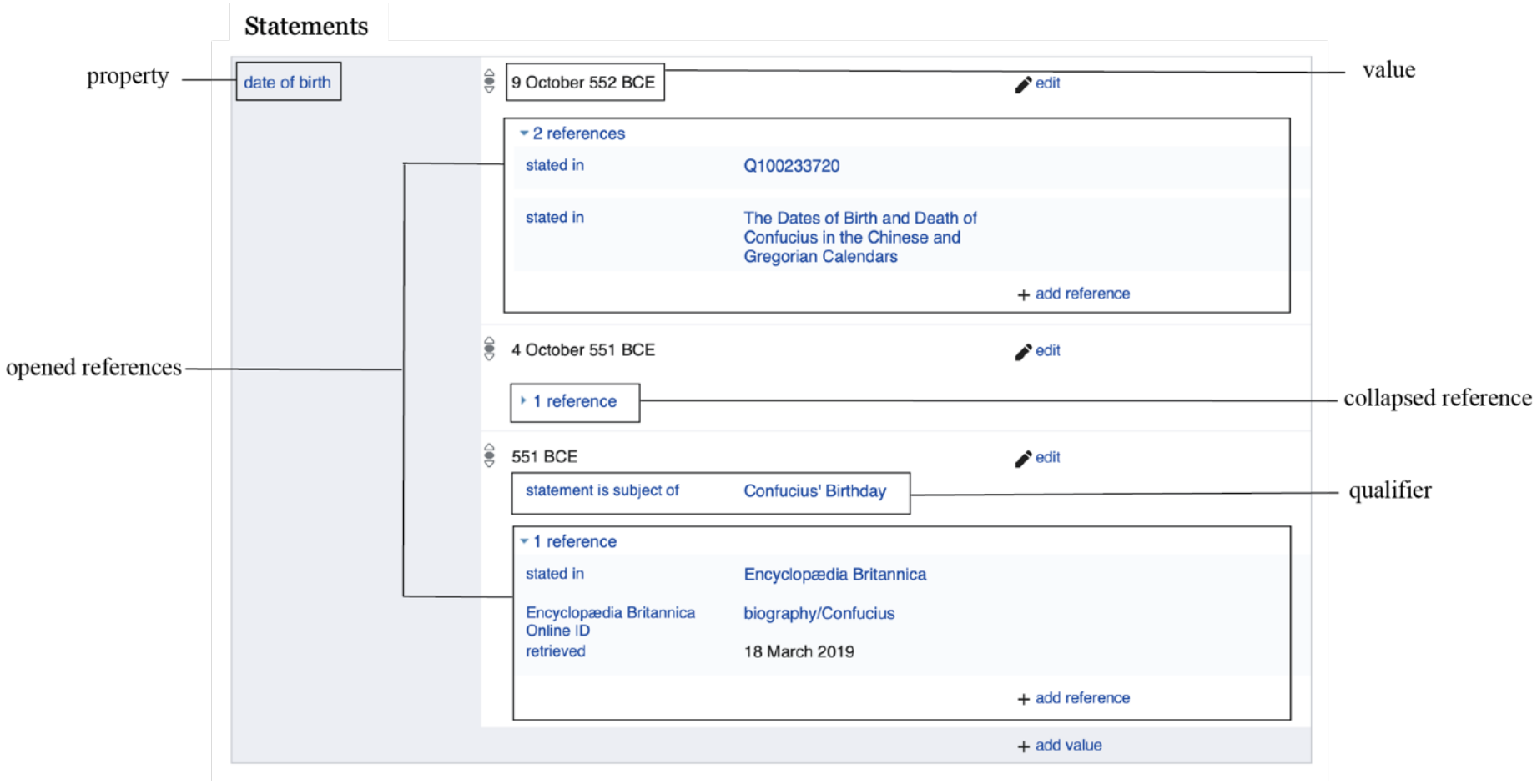
On the interface, an item is displayed in **4** parts:

1) **basic information** includes label, item identifier, description, aliases, and their multilingual displays.
2) **statements**, which describe features of an item, consists of property and value.
3) **external identifiers** link an item *to* external databases.
4) **sitelinks** connect an item to other Wikimedia projects, like Wikipedia.

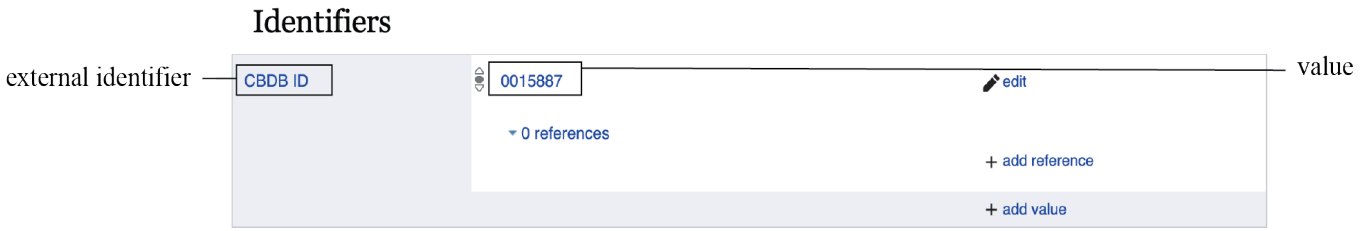
Wikidata and its data model (by looking at a modeled instance)

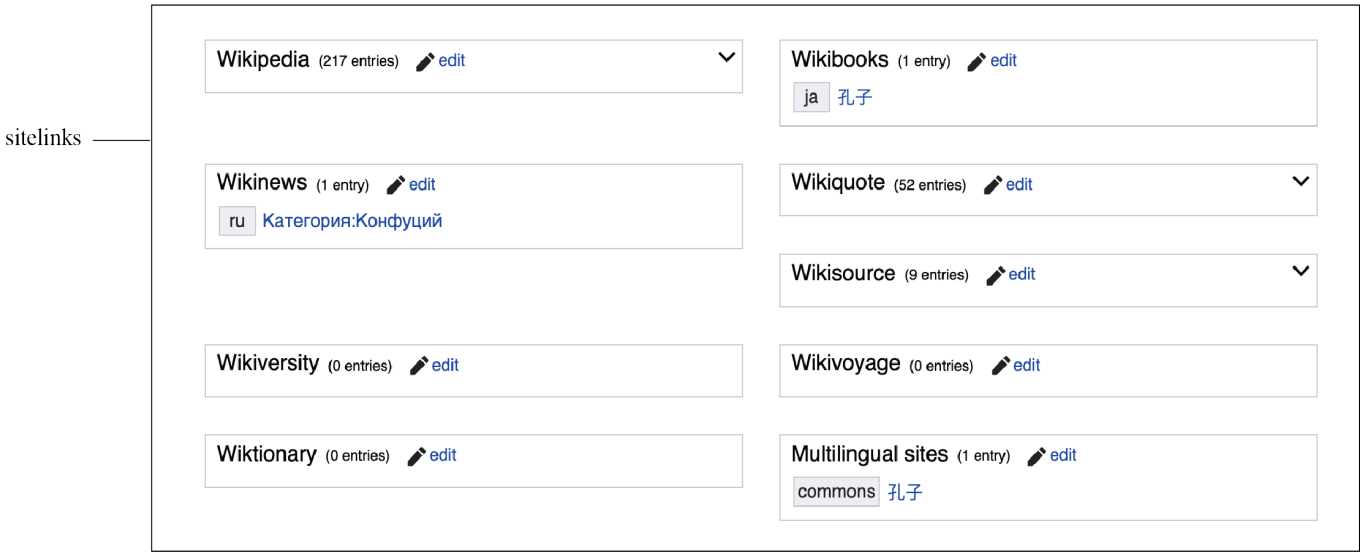


Wikidata and its data model

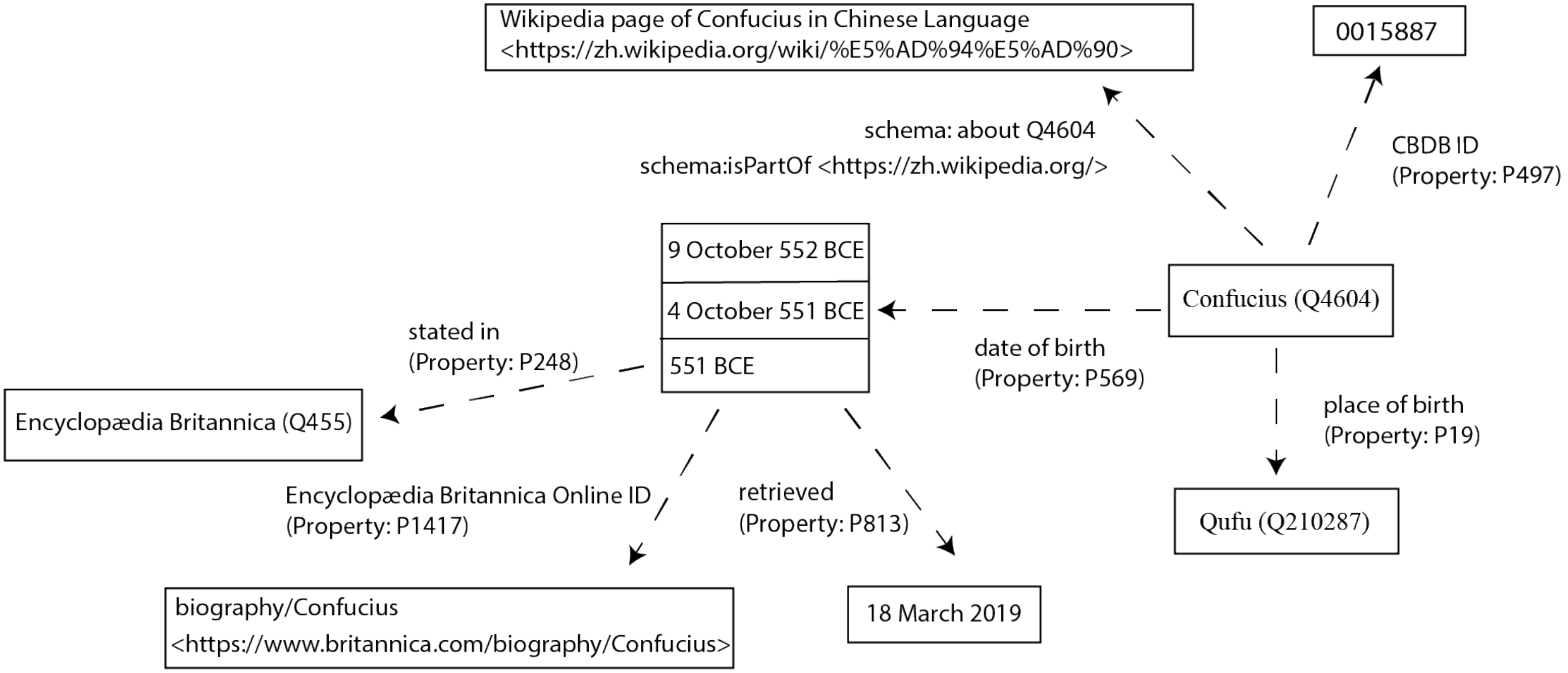


Wikidata and its data model





Wikidata and its data model



2. Some cases in Chinese studies

2. Some cases in Chinese studies

people
https://www.wikidata.org/wiki/Q45681209
https://www.wikidata.org/wiki/Q4604

work
https://www.wikidata.org/wiki/Q10883609

encyclopedia
https://www.wikidata.org/wiki/Q699477

External identifiers on Wikidata

- Academia Sinica authority ID
- CBDB ID
- ctext data entity ID

V · T · E	China (Q29520)-related properties
Language-related (Chinese)	pinyin transliteration · Unicode code point · stroke count · radical · residual stroke count · Han character in this lexeme · CJKV variant character · GlyphWiki ID · fanqie · Jyutping transliteration · Yale romanization · Cantonese Pinyin · Cantonese Transliteration Scheme transliteration
Language-related (others)	Wylie transliteration · Tibetan pinyin · THL Simplified Phonetic Transcription · Möllendorff transliteration
Culture-related	ancestral home · dan/kyu rank · Eight Banner register · courtesy name · temple name · posthumous name · art-name · era name
Identifiers (for persons)	CALIS ID · NLC authorities · Shanghai Library person ID · Academia Sinica authority ID · CBDB ID · **ctext data entity ID** · Dharma Drum Institute of Liberal Arts person ID · Douban author ID · Douban Read author ID · China Vitae person ID · Chinese Political Elites Database ID · BDRC Resource ID · Sina Weibo user ID · Douban username · Zhihu username · Bilibili user ID · QQ user ID · TikTok username · Douyin ID · Chinese Olympic Committee athlete ID · ChinesePosters artist ID · Douban movie celebrity ID · Mtime people ID · Douban musician ID · QQ Music singer ID · NetEase Music artist ID · Baidu ScholarID · Arnet Miner author ID · China Engineering Expert Tank ID · CNKI author ID · bgm.tv person ID · China Martyrs ID · Dictionary of Anhui Writers ID · 20th Century Chinese Biographical Database ID · Biographical Dictionary of Chinese Christianity ID · TBDB ID · 1905.com star ID
	HKMDB person ID · Webb-site person ID · DLCM ID · HKCAN ID · Moov artist ID
Identifiers (others)	China administrative division code · CHGIS ID · Dharma Drum Institute of Liberal Arts place ID · Fuzhou Architecture Heritage ID · Shanghai Library place ID · Code for China Reservoir Name · China railway TMIS station code · Chinese Library Classification · CN · Unified book number · Flora of China ID · IECIC 2015 ID · ctext work ID · Douban book version/edition ID · Douban Read eBook ID · Douban book series ID · Douban book works ID · NLC Bibliography ID (Chinese language) · NLC Bibliography ID (foreign-language) · Douban film ID · Mtime movie ID · 1905.com film ID · Douban drama ID · Douban game ID · TapTap application ID · A9VG game ID · Douban music ID · Douban site name · bgm.tv subject ID · bgm.tv character ID · Moegirlpedia ID · HuijiWiki wiki ID · Bilibili bangumi ID · Bilibili tag ID · Bilibili video ID · Douyin video ID · TikTok music ID · CJFD journal article ID · CNKI CJFD journal ID · Baidu Scholar paper ID · Baidu Scholar journal ID · Arnet Miner publication ID · CQVIP article ID · Xikao History ID · Xikao Repertoire ID · Chinese Clinical Trial Registry ID · China Treaty Database ID · National Database of Laws and Regulations ID · PKULaw CLI Code · AppGallery app ID · Chinese Painting Database ID · Zhihu topic ID · Encyclopedia of China ID (Second Edition) · Encyclopedia of China ID (Third Edition) · **ctext data entity ID** · WeChat ID · Baijiahao ID · Internet Content Provider Registration Record ID · Unified Social Credit Identifier · Qichacha firm ID · ARWU university ID · Chinese School Identifier · QQ Music album ID · QQ Music track ID · CNGAL entry ID · HuijiWiki article ID · CNKI institute ID · Hmoegirl ID · Gitee username · TGbus ID · TGbus franchise ID
	HKMDB film ID · Hong Kong film rating
Open proposals	None at present
See also: {{Wikidata Han character properties}} – {{Taiwan properties}}	

Measure the effectiveness of being on WD

- CBDB as an example

https://www.wikidata.org/wiki/Wikidata:WikiProject_East_Asia/China_Biographical_Database_import

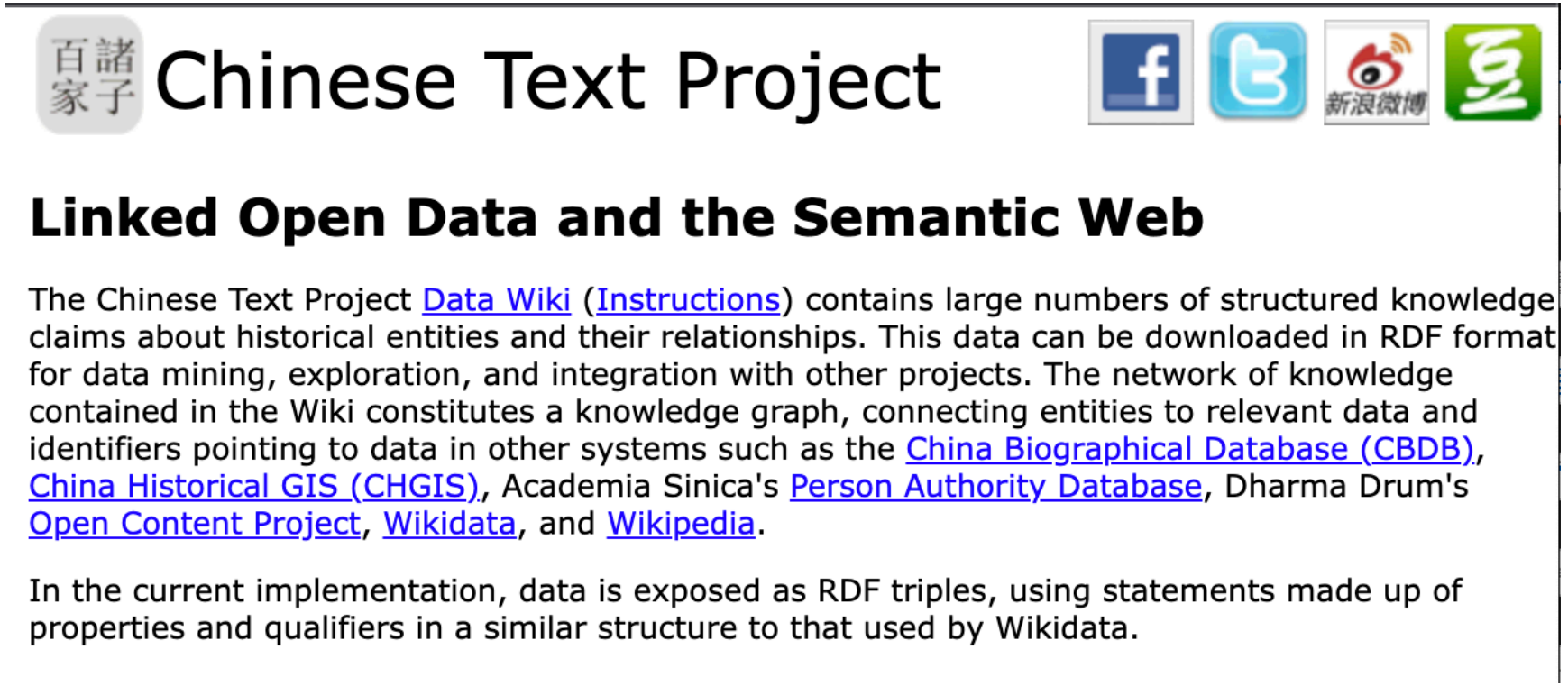
Wikidata identifiers in projects

- Buddhist Studies Person Authority Databases

https://sdp.chibs.edu.tw/person/index.php?fromInner=A003970

Linked Open Data and the Semantic Web for Chinese studies

- https://ctext.org/tools/linked-open-data



How to be on Wikidata

- https://www.wikidata.org/wiki/Property:P9613
- https://www.wikidata.org/wiki/Property_talk:P9613
- https://www.wikidata.org/wiki/Wikidata:Property_proposal/Authority_control
- https://www.wikidata.org/wiki/Wikidata:External_identifiers

3. Hands-on session

3. Hands-on session

Try curating one or two items on Wikidata following the workflow:

1. Choose one item (organisation, people, place, concept, work etc.) found in TLS (https://hxwd.org/index.html)/ Kanseki Repository (https://www.kanripo.org)
2. Find the corresponding Wikidata entity of the chosen item or create one entity if it does not have one.
3. Curate this Wikidata entity using the reliable resources you have found.
4. Try querying, downloading, and visualising the curated entity on Wikidata.
5. Compare and connect the Wikidata entity with that of the original item and reflect on the process.

Resources

- Wikidata data model

https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer

- Wikidata List of Properties (a starting point)

https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/all

- (VISUALISE) Wikidata Graph Builder

https://angryloki.github.io/wikidata-graph-builder/

- (QUERY) Wikidata SPARQL examples

https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/queries/examples

- (DOWNLOAD) Wikidata Data access

https://www.wikidata.org/wiki/Wikidata:Data_access

Wikidata:
Lexicographical data/linguistic data

- https://www.wikidata.org/wiki/Lexeme:L521650
- https://www.wikidata.org/wiki/Wikidata:Lexicographical_data

4. Wikidata for a comprehensive collaborative research environment for the study of premodern Chinese culture?

5. A review of Wikidata in DH projects

Research Questions

- **Q1**: How is Wikidata described in the current DH literature?
- **Q2**: To what end is Wikidata being experimented within the DH domain?
- **Q3**: What are the potentials of embracing Wikidata in data-related activities in DH projects?
- **Q4**: What are the challenges and possible solutions in DH projects regarding Wikidata's data quality?

Methodology

- a systematic review (Kitchenham, 2004) of DH-related projects which adopted Wikidata

ID	Title	References	Description	Q1	Q2	Q3	Q4
p1	Corpus-DB: a Scriptable Textual Corpus Database for Cultural Analytics	Reeve (2020)	Corpus-DB (corpus-db.org) is a database and query framework which solves the problems of text retrieval, text cleaning, corpus compilation, and metadata aggregation that often form the first step for researchers in computational text analysis. Wikidata is a source of metadata to enrich the corpus in the process of metadata aggregation.	generic, content provider	metadata curation	data consumption	-

Q1 How is Wikidata described in the current DH literature?

Technology Stack		Platform	Content	
web technologies including URI, HTTP, Unicode, etc.	Linked Data	dissemination platform	Content	Form
			multilingual open free generic editable heterogeneous global online	datasets (i.e. linked open data, authority file, controlled vocabulary)
				database
RDF dumps		Wikimedia platform		ontology
live SPARQL endpoint		linking hub		knowledge base (ontology + instances
				knowledge graph

Q2 To what end is Wikidata being experimented within the DH domain?

- Annotation and enrichment (text corpus, linguistic dictionaries, datasets, journal articles, annotation tool)
- Metadata Curation (integration and interoperability of authority datasets, improvement of metadata quality by sharing their metadata for public use or using Wikidata information to enhance local records)
- Modelling (knowledge modelling in the context of the Semantic Web, reference for the creation of data models)
- NER-related tasks and Miscellaneous (pedagogy)

Q3 What are the potentials of embracing Wikidata in data-related activities in DH projects?

- Data Consumption (40)
  - Wikidata as a content provider is better explored
  - annotation, metadata curation, modelling, NER, pedagogy
- Data Publication and Exchange (11)
  - Potential lies in Wikidata as a platform and a technology stack
  - data integration (a linking hub for domain resources)
  - data production (a low-tech approach to publish linked data)
  - production for consumption

Q4 What are the challenges and possible solutions associated with Wikidata in DH projects?

- Challenges

technical challenges (identifier mismatches, data model incompatibility etc.), concern about its quality due to open model (data coverage > data accuracy)

- What can we learn from the reviewed projects?

evaluate its quality against other available domain sources, form a community of practice in your own field, design a workflow that orchestrate technical and labour resources from the projects and Wikidata.

Conclusion

- More than a knowledge base: a content provider, a platform, a technology stack.
- mainly used for: annotating research materials, curating metadata, publishing and interlinking linked data in knowledge modelling.
- Consumption > Publication. More potential in taking it as a platform and a technology stack to create an ecosystem of data sharing.
- take into consideration: available domain sources, community practices, workflow design that balance resources from the projects and Wikidata.

References

- **Cook, S.** (2017). The uses of Wikidata for galleries, libraries, archives and museums and its place in the digital humanities. *Comma*, **2017**(2):117-124.
- **Kitchenham, B.** (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, **33**(2004), 1-26.
- **Mora-Cantallops, M., Sánchez-Alonso, S. and García-Barriocanal, E.** (2019). A systematic literature review on Wikidata. *Data Technologies and Applications*, **53**(3): 250–68.
- **Tharani, K.** (2021). Much more than a mere technology: A systematic review of Wikidata in libraries. *The Journal of Academic Librarianship*, **47**(2).

Back to the question:
Wikidata for a comprehensive collaborative research environment for the study of premodern Chinese culture?