Durham University

Text and Data Mining with the Chinese Text Project (ctext.org)

Online materials for this session:
https://dsturgeon.net/dsea2024

Donald Sturgeon
Department of Computer Science
Durham University, UK
donald.j.sturgeon@durham.ac.uk

Overview of this session

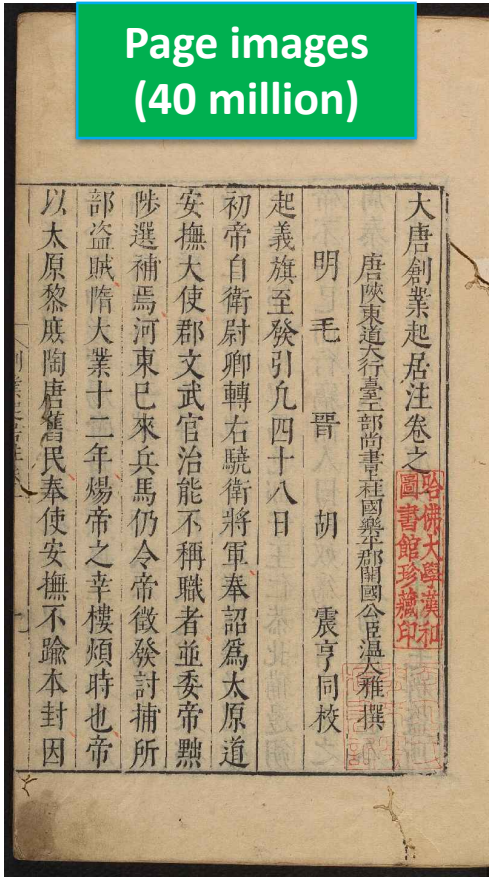
- General introduction to ctext.org
- Locating and using texts on ctext.org
  - Searching
  - Editing
- Digital tools for textual analysis & visualization
  - Text reuse
  - Pattern search (regular expressions)
  - Visualization
- Historical data: knowledge graphs and semantic annotation

Chinese Text Project (ctext.org), 2005-present

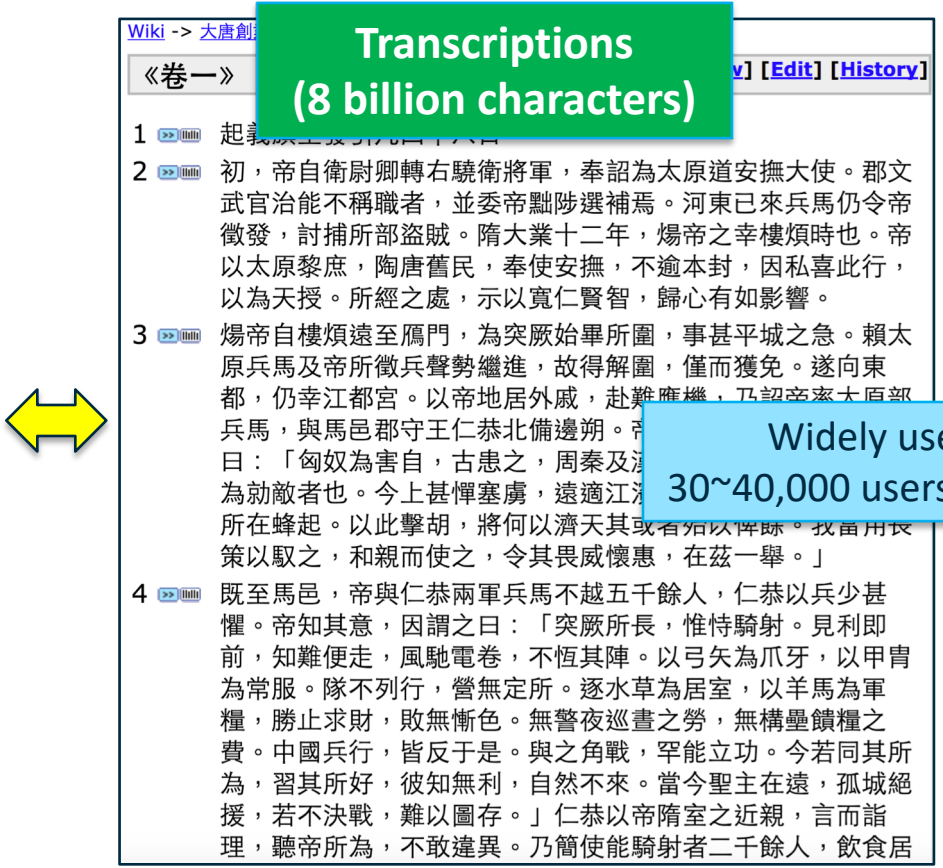
- A digital library and full-text database of premodern Chinese sources
- Main goals:
  - Organize premodern Chinese-language primary sources digitally
  - Create digital texts that are more than just surrogates for print
  - Do this in a way that scales to large volumes of material
    - Use automation where possible
    - Use crowdsourcing to correct and improve on imperfect digitization

ctext.org – Main types of material

Primarily woodblock printed and handwritten sources in classical Chinese



Page images (40 million)



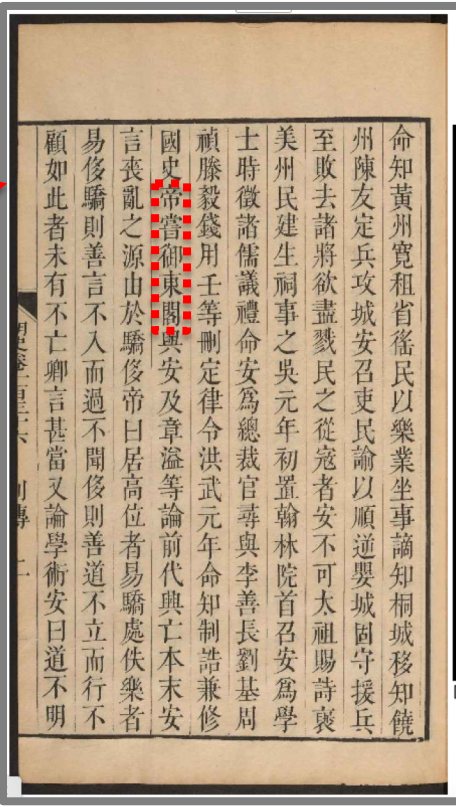
Transcriptions (8 billion characters)

Widely used: 30~40,000 users per day

Image sequences and literal transcriptions

《列傳第二十

5 洪武元年
代興亡
驕，處
而行不
「道不
美色之
曰：「
舊。及
士，翰
「安寧
之。



命知黃州寬租省徭民以樂業坐事謫知桐城移知饒
州陳友定兵攻城安召吏民諭以順逆嬰城固守援兵
至敗去諸將欲盡戮民之從寇者安不可太祖賜詩褒
美州民建生祠事之吳元年初置翰林院首召安為學
士時徵諸儒議禮命安為總裁官尋與李善長劉基周
禎滕毅錢用壬等刪定律令洪武元年命知制誥兼修
國史帝嘗御東閣與安及章溢等論前代興亡本末安
言喪亂之源由於驕侈帝曰居高位者易驕處佚樂者
易侈驕則善言不入而過不聞侈則善道不立而行不
顧如此者未有不亡卿言甚當又論學術安曰道不明

命知黃州寬租省徭民以樂業坐事謫知桐城移知饒
州陳友定兵攻城安召吏民諭以順逆嬰城固守援兵
至敗去諸將欲盡戮民之從寇者安不可太祖賜詩褒
美州民建生祠事之吳元年初置翰林院首召安為學
士時徵諸儒議禮命安為總裁官尋與李善長劉基周
禎滕毅錢用壬等刪定律令洪武元年命知制誥兼修
國史帝嘗御東閣與安及章溢等論前代興亡本末安
言喪亂之源由於驕侈帝曰居高位者易驕處佚樂者
易侈驕則善言不入而過不聞侈則善道不立而行不
顧如此者未有不亡卿言甚當又論學術安曰道不明

[View] [Edit] [Quick edit] [Editing help]

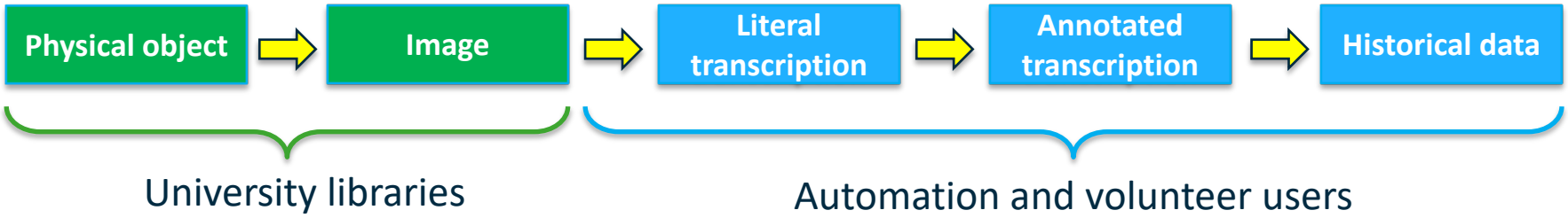
Item
Scanr
Full t
明史

1. 本紀第一　太祖一
2. 本紀第二　太祖二
3. 本紀第三　太祖三
4. 本紀第四　恭閔帝
5. 本紀第五　成祖一
6. 本紀第六　成祖二
7. 本紀第七　成祖三
8. 本紀第八　仁宗
9. 本紀第九　宣宗
10. 本紀第十　英宗前紀
11. 本紀第十一　景帝
12. 本紀第十二　英宗后紀
13. 本紀第十三　憲宗一
14. 本紀第十四　憲宗二
15. 本紀第十五　孝宗
16. 本紀第十六　武宗
17. 本紀第十七　世宗一
18. 本紀第十八　世宗二
19. 本紀第十九　穆宗
20. 本紀第二十　神宗一
21. 本紀第二十一　神宗二

Durham University

Chinese Text Project (ctext.org), 2005-present

- OCR applied to all materials to generate transcriptions & facilitate search
- Crowdsourced editing interface to correct mistakes in transcriptions
- Crowdsourced annotation interface to enrich texts and create data
- API and tools for text mining of a large collection of material

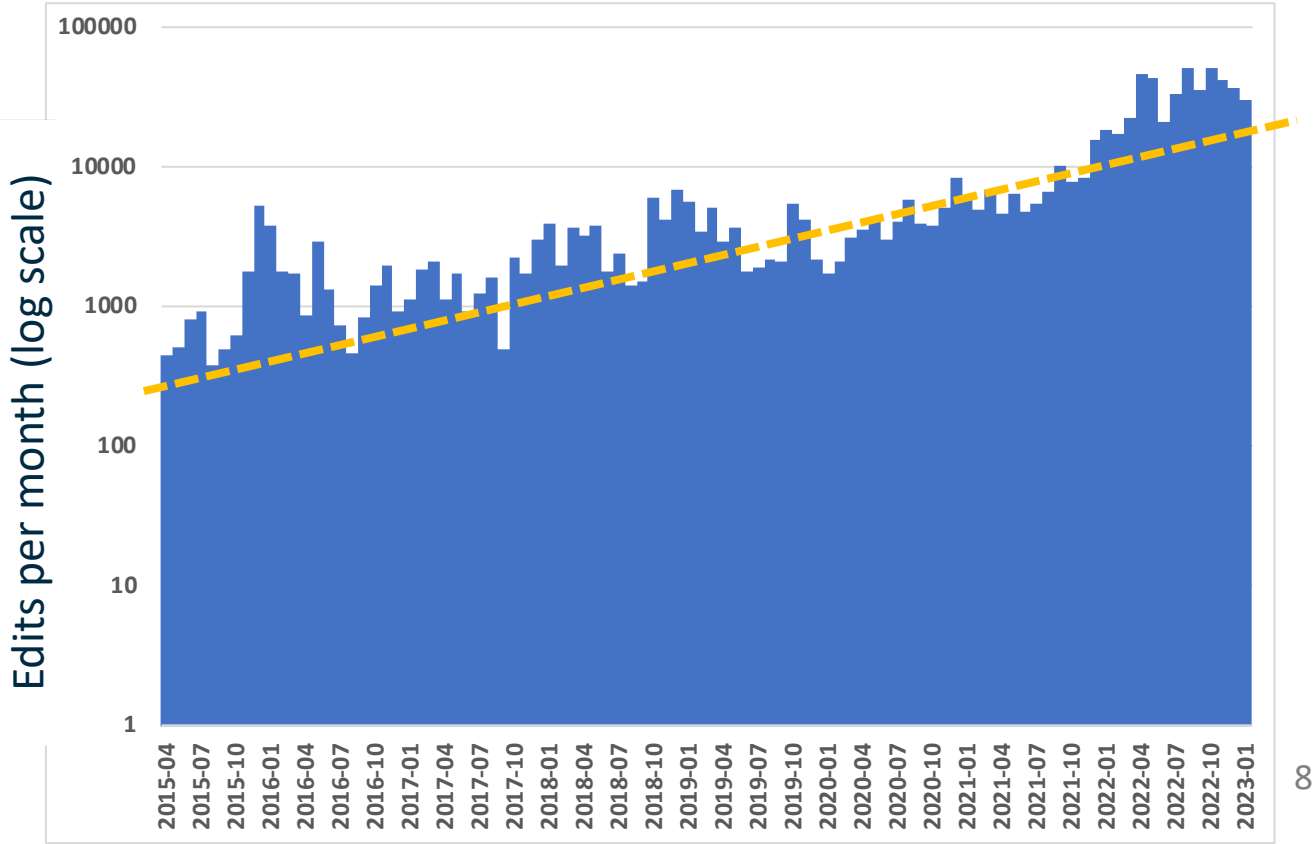


Important: ctext is not a “traditional” text database

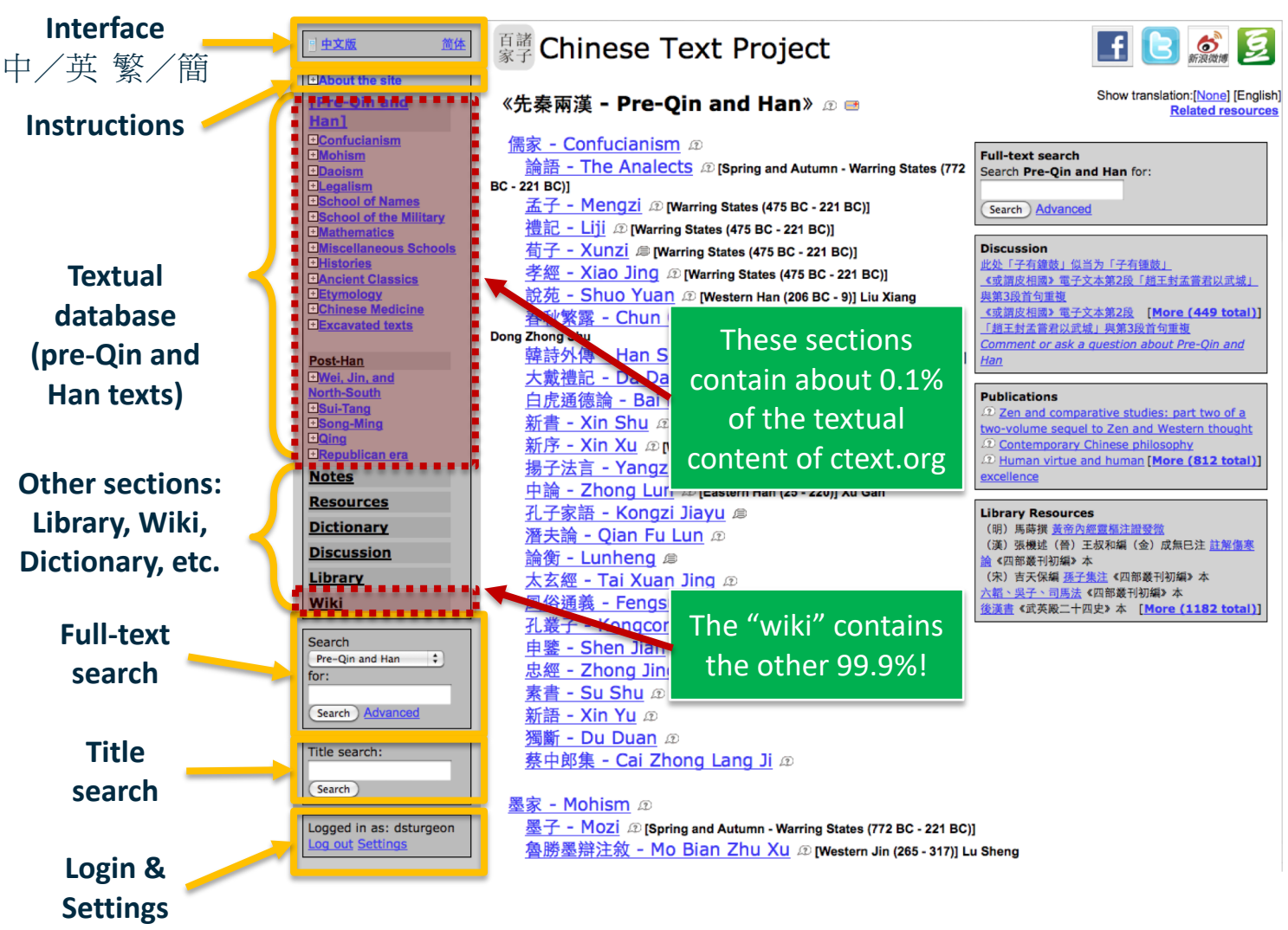
- Not a collection of reviewed, authoritative text
- Databases of this type:
  - Academia Sinica 漢籍全文
  - CHANT / 漢達文庫, etc.
- Instead: methods of navigating primary sources
  - Authority does not derive from expert review
    - Instead: verification of evidence by individual users
      - Transcriptions of primary sources
      - Data/claims as evidenced by primary sources

People contribute to ctext

Exponential growth in edits since use of crowdsourcing



Durham University



Durham University

Acknowledgments

HARVARD LIBRARY

Princeton University LIBRARY

香港中文大學圖書館
The Chinese University of Hong Kong Library

CBDB

China Historical GIS

法鼓文理學院
DILA Dharma Drum Institute of Liberal Arts

ctext.org users

Hands-on tutorial: Part I
Basic use of ctext.org

- Overview
  - Setup
  - Finding texts, searching in texts, locating in scans
  - Special functions in the textual database
    - Parallels, translations, commentary
- Editing
- Plugins
- (Tutorial: “Practical introduction to ctext.org”)

Finding Texts

Left-hand side => “Title search”

Possible results:

Transcription (text DB)
(not user editable)

Transcription (wiki)
(user editable)

Transcription (OCR, wiki)
(uncorrected, editable)

Scanned primary source
(not a transcription)

Example:

Indicates this transcription is *linked* to a scanned representation of the 四庫全書 edition of the text

Durham University

Hands-on tutorial: Part II

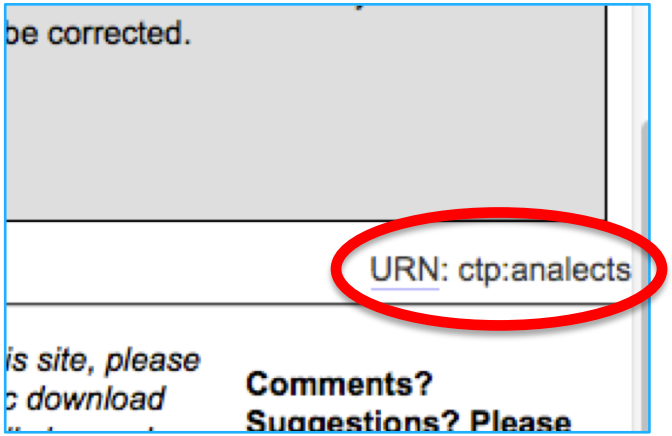
Text mining with ctext's Text Tools plugin

- Textual analysis tools
  - N-gram counts
  - Text reuse identification using n-grams
  - Regular expressions
  - Cosine similarity
  - Principal Component Analysis
- Visualization tools
  - Network graphs
  - Heat maps
  - Charts
- (Tutorial: "Text Tools for ctext.org")

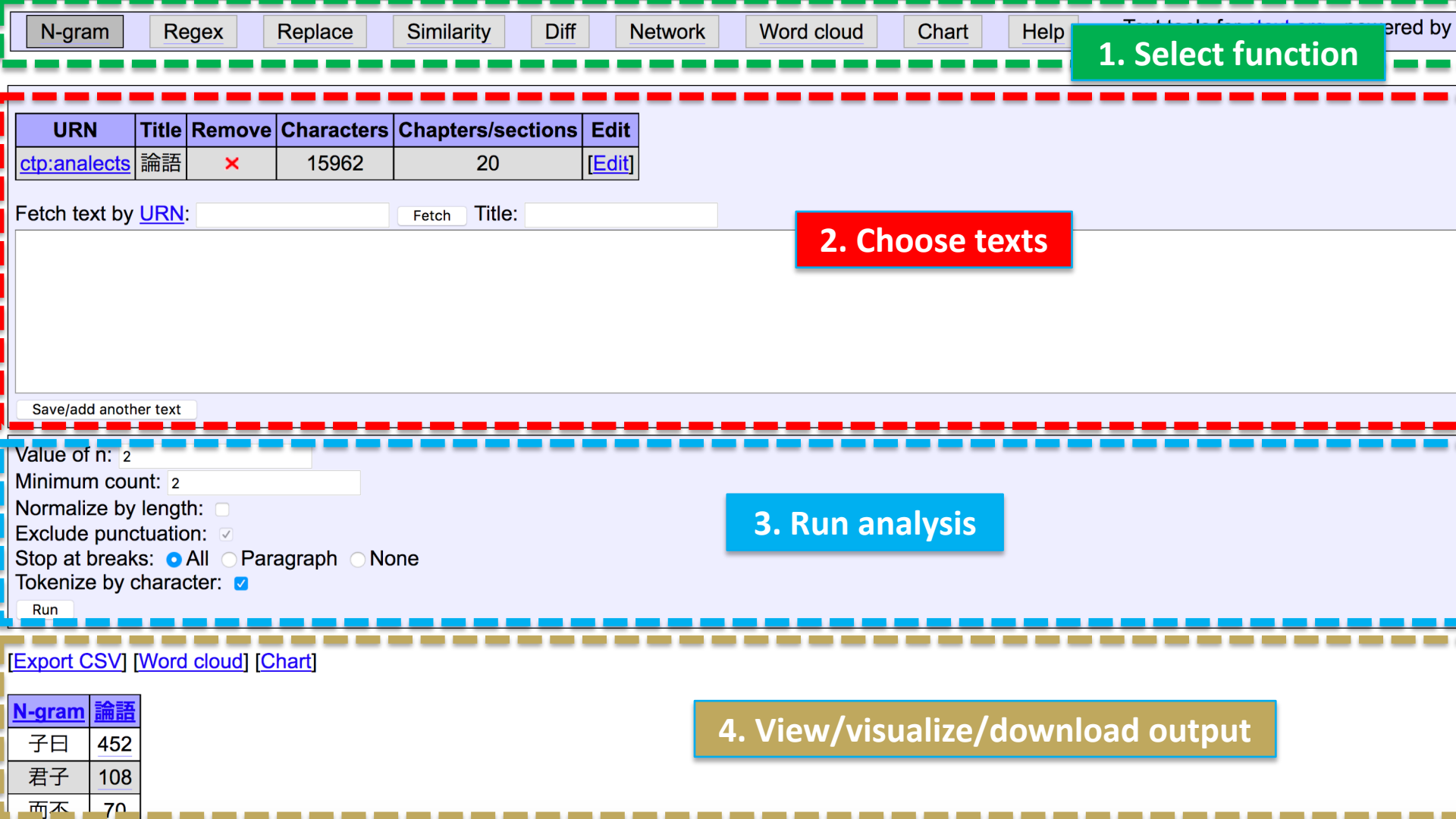
CTP URNs

Digital identifiers for specific textual items

- URNs identify textual objects
- To find: open the contents page for the text
  - Look at bottom-right corner
  - CTP URNs always begin “ctp:…”



- To “decode”: same as finding texts by title:
  - Paste URN into “Title search box”
  - Click “Search”
  - Contents page for that text will open


N-gram
Regex
Replace
Similarity
Diff
Network
Word cloud
Chart
Help
1. Select function
URN
Title
Remove
Characters
Chapters/sections
Edit
ctp:analects
論語
15962
20
[Edit]
Fetch text by URN:
Fetch
Title:
2. Choose texts
Save/add another text
Value of n: 2
Minimum count: 2
Normalize by length:
Exclude punctuation:
Stop at breaks: All Paragraph None
Tokenize by character:
Run
3. Run analysis
[Export CSV] [Word cloud] [Chart]
N-gram
論語
子曰
452
君子
108
4. View/visualize/download output

Hands-on tutorial: Part III
Data and semantic annotation

- Goals:
  - Make semantic elements of texts machine-readable
  - Explicitly record historical data contained in premodern works
    - E.g. biographical, geographical, bureaucratic, bibliographic, …
- Currently most relevant to historical texts
  - E.g. the 25 standard histories
    - All of these currently have at annotations & associated data
    - These have not yet been comprehensively annotated
    - N.B. Only one edition of each work is chosen to be annotated

Conceptualization of data

In ctext, “data” consists of:

1. Annotations of text connecting text to entities
    - Only contain the most basic information:
        - Entity type (person / place / work / …)
        - Entity identifier
        - For dates only: year, month, day of the date (in Chinese terms)
2. Knowledge claims about entities
    - Subject – verb – object
        - Optionally, “qualifiers” as adverb – object pairs

Adding explicit semantic information to a text

安石榴不可多食，損人肺。

One must not eat too many pomegranates, as it harms the lungs.

辛亥，以安石為尚書左僕射兼門下侍郎。

[On day] 48,	[the emperor] made Anshi vice director…

Context: dynastic narrative record, in which Wang Anshi has just been mentioned.

Which day? → 5 August 1075 AD (Julian)

Which "Anshi"? → Wang Anshi, well-known politician

後以疾卒。著《兵略》，世頗稱之。子安石。

Later [he] died of illness.	[He] authored *Bing Lue*, which was widely praised.	[His] son was Anshi.

Context: biography of Chen Guan in the same work.

Which Bing Lue? → The *Bing Lue* that was written by Chen Guan

Which "Anshi"? → Chen Anshi, son of Chen Guan

Durham University

Creating annotations requires domain knowledge

There are easy cases (famous individuals, with no same-named individuals)

Many cases are much harder!

Name	Type
王佐	person
王佐	person
王佐	person
王佐	person
王佐	person
王佐	person
王佐	person
王佐	person
王佐	person
王佐	person
王佐	person
王佐	person
王佐	person
王佐	person
王佐	person
王佐	person

Efficient annotation of entities *requires* having detailed data about those entities

Chinese Wikipedia disambiguation page for “王佐”

宋朝 [編輯]

- 王佐 (宋朝)，山陰（今浙江紹興）人，字宣子，號敬齋。紹興十八年（1148）戊辰科狀元。

明朝 [編輯]

- 王佐 (杭州通判)，明朝洪武年間杭州府通判。
- 王佐 (洪武進士)，明朝洪武二十一年進士。
- 王佐 (戶部尚書)，明朝永樂九年舉人，官至戶部尚書。
- 王佐 (永樂進士)，明朝永樂十九年進士。
- 王佐 (宣德進士)，明朝宣德二年進士。
- 王佐 (正統舉人)，明朝正統十二年舉人。
- 王佐 (景泰進士)，明朝景泰二年進士。
- 王佐 (天順進士)，明朝天順元年進士。
- 王佐 (成化河南進士)，河南南陽府汝州人，明朝成化八年進士。
- 王佐 (成化直隸進士)，北直隸開州人，明朝成化八年進士。
- 王佐 (南京戶部尚書)，明朝成化十四年進士，官至南京戶部尚書。
- 王佐 (嘉靖舉人)，福建同安人，明朝嘉靖元年舉人。
- 王佐 (工部尚書)，明朝萬曆十一年進士，官至工部尚書。
- 王佐 (萬曆進士)，明朝萬曆十四年進士，湖廣武陵人
- 王佐 (崇禎舉人)，湖廣河陽人，明朝崇禎三年舉人。
- 王佐 (崇禎進士)，明朝崇禎四年進士。

清朝 [編輯]

- 王佐 (雍正進士)，清朝雍正元年進士。
- 王佐 (乾隆進士)，清朝乾隆十九年進士。
- 王佐 (資政院)（1853-1931），字寄廎，上虞豐惠人。光緒十五年（1889）恩科舉人。主修《上虞縣誌》。

Entity records and knowledge claims

王珪 [View] [Ed
ctext:533428

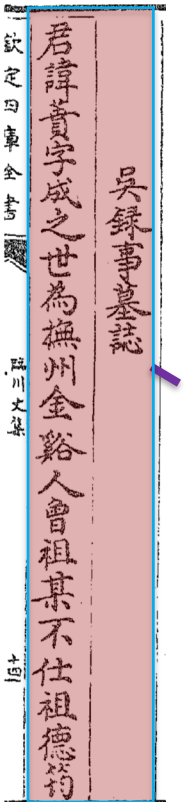
Entity ID (subject)

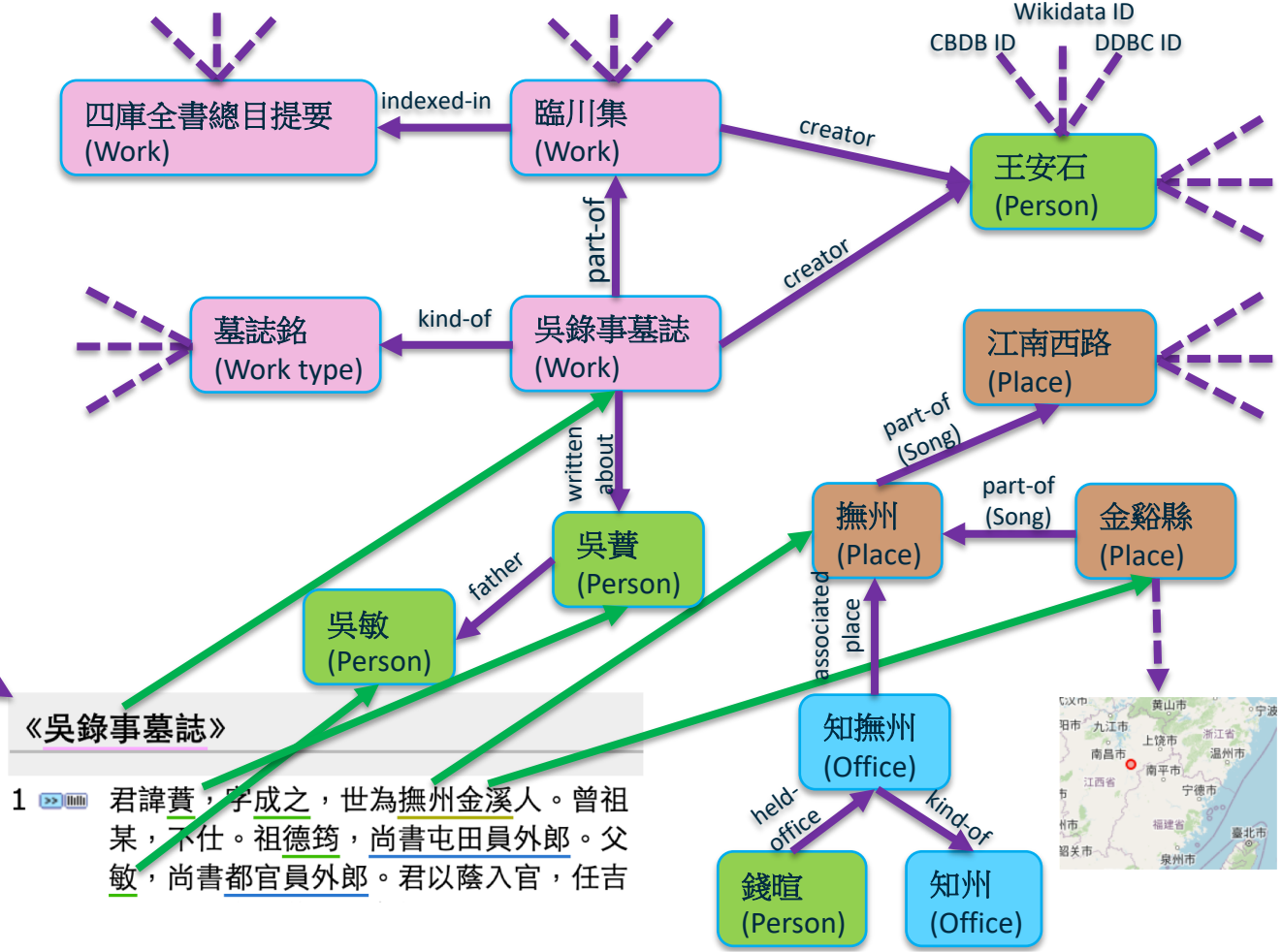
Other same-named entities

See also: 王珪 (ctext:706573) 王珪 (ctext:629741)

Verb	Object	Textual evidence
type	person	
name	王珪	
name-style	禹玉	《宋史·列傳第七十一》：王珪，字禹玉，成都華陽人，後徙舒。
associated-place	place:華陽縣	《宋史·列傳第七十一》：王珪，字禹玉，成都華陽人，後徙舒。
born-date	天禧己未年七月九日 1019/8/12	《文昌雜錄·第六卷》：左僕射王公珪，己未七月初九日生。
died-date	元豐八年五月庚戌 1085/6/12	《宋史·本紀第十七》：庚戌，王珪薨。
authority-cbdb	1845	Authority identifiers
authority-wikidata	Q45359570	
link-wikipedia_zh	王珪_(宋朝宰相)	Links to the entity in Wikipedia
link-wikipedia_en	Wang_Gui_(Song_dynasty)	
held-office	office:參知政事	
from-date	熙寧三年十二月丁卯 1071/1/14	《宋史·本紀第十五》：丁卯，以韓絳、王安石並同中書門下平章事，王珪參
held-office	office:山陸使	

Semantic annotation & knowledge graph





Data: basic searching

- Go to the “Data wiki”
- Type your query in the “Data search” box
- Simple queries (e.g. a search term) will find cases where:
  - <*some entity*> <**name**> <*your-query-text*>
  - E.g. a query for “崇寧” will find things that can be named “崇寧”, e.g.
    - The Song dynasty era 崇寧
    - The place 崇寧縣

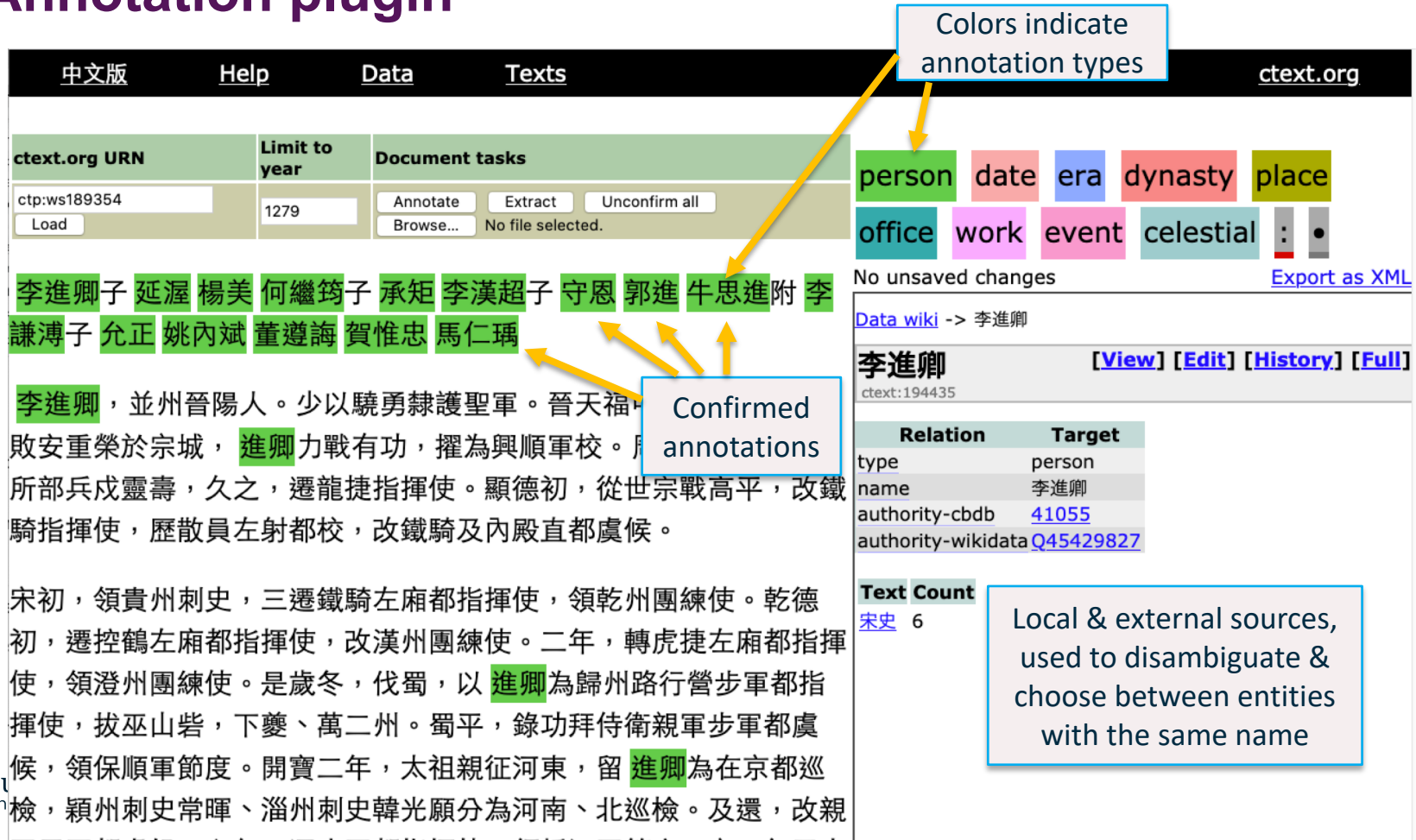
Data: basic searching

- More general syntax:
  - property:value
    - E.g. name:至德
      - Matches any entity with this name
    - E.g. held-office:樞密使
- % is a wildcard
  - E.g. name:趙%
    - Matches 趙 (dynasty), 趙王 (office), 趙禎=宋仁宗 (person), …

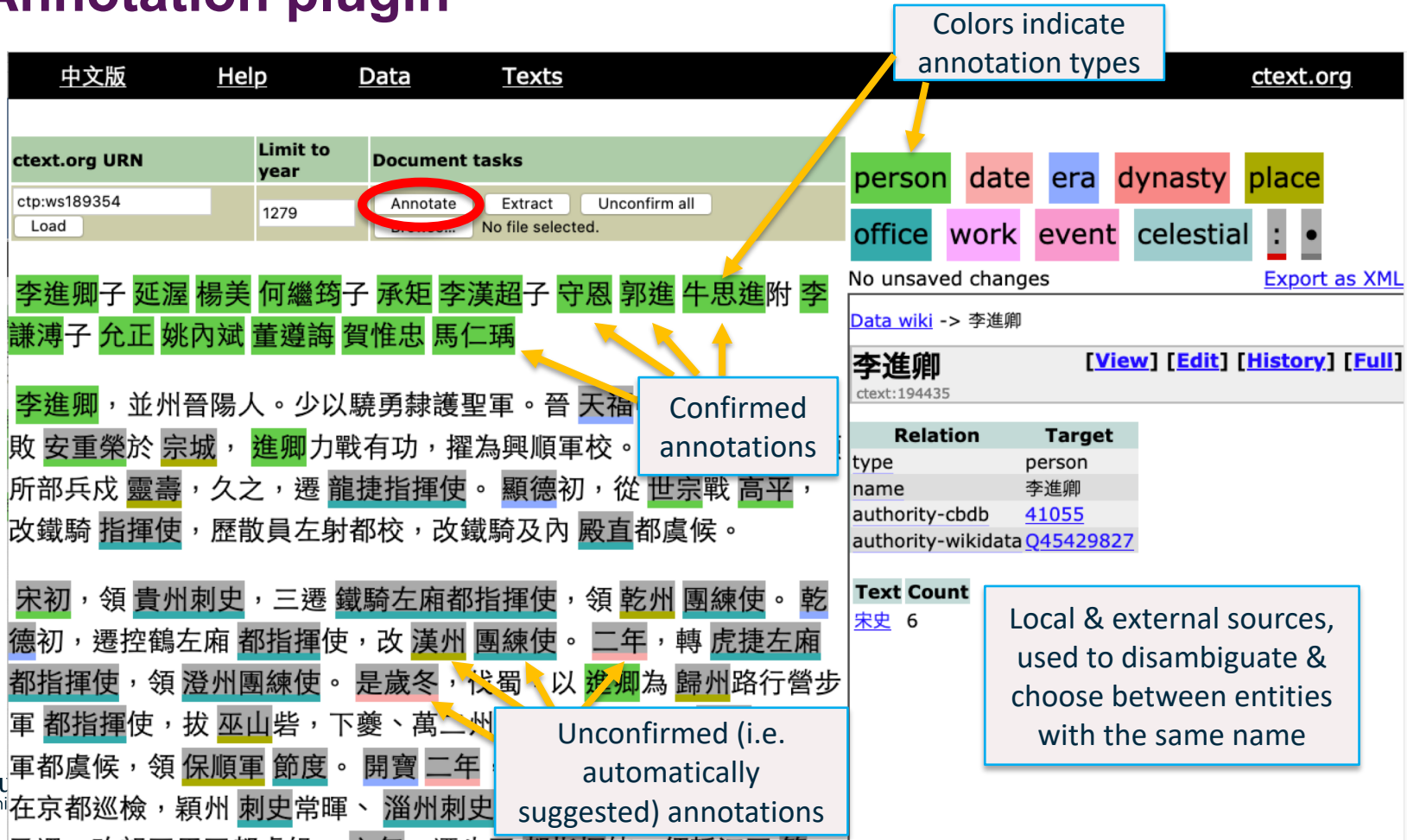
Data: basic searching

- Objects can be specified by name or entity ID
  - E.g. held-office:樞密使 vs held-office:ctext:85216
  - Easier to use the name, assuming that it is unambiguous
- The data wiki itself offers suggested searches
  - Look for an example containing a similar claim
    - Usually will be a link generating a search specification
- A URN matches all entities annotated in a text
  - E.g. ctp:wb975976 matches entities that occur in the text of the 宋史
- Space-separated clauses are conjunctive
  - E.g. name:趙% type:person ctp:wb975976
    - Matches all people surnamed 趙 referenced in the 宋史

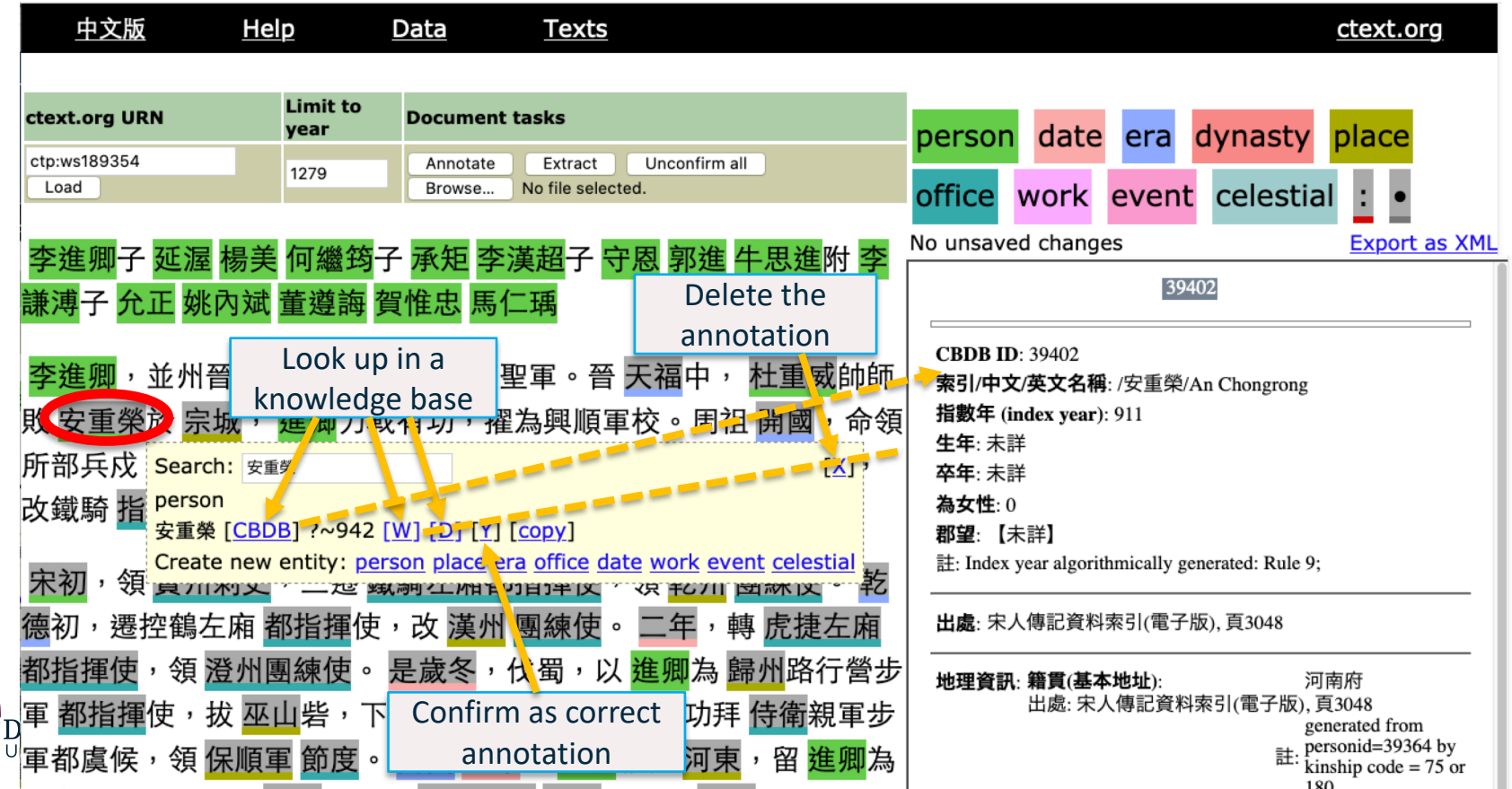
Annotation plugin



Annotation plugin



Annotation plugin



Annotation plugin

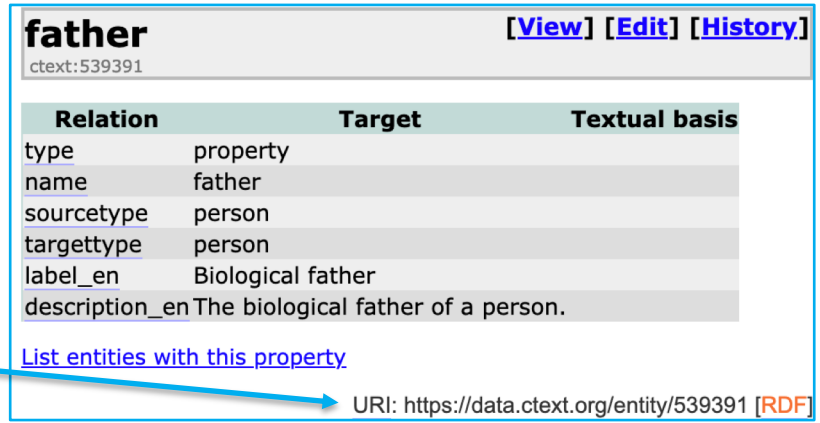
Durham University

Data: editing principles

- Standards of evidence:
  - For textual content edits, evidence is the scan
    - Transcriptions are always based on one edition
    - Markup can be used to highlight errors in the text itself
  - For knowledge claims, evidence is a line of text
    - E.g. “Zhu Xi died on the date 慶元六年三月甲子”
      - Evidence: 《宋史·本紀第三十七》:三月甲子，朱熹卒。
        - Recorded as a machine-readable citation
  - It follows that many true things are not included!
    - Goal: machine-readable, grounded & transparent dataset

RDF data model: assigning globally unique identifiers

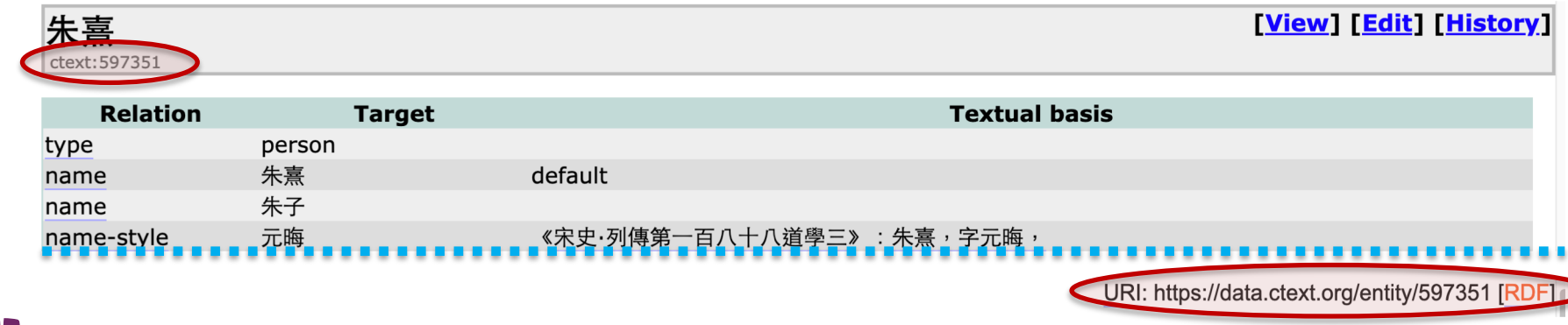
- Goal: combined querying of data from multiple sources (databases)
- Problems
  - Our identifiers may be different (“person-name” vs “name” vs …)
  - Our identifiers may be the same but mean different things!
- Solution: use URIs (syntactically identical to URLs) to identify *concepts*
- E.g. the historical person named “朱熹”: https://data.ctext.org/entity/597351
- E.g. the concept of biological father: https://data.ctext.org/entity/539391
  - Shown in the Data Wiki just like URNs:



Durham University

RDF data model: assigning globally unique identifiers

- To make things easier to read, we abbreviate the URIs
  - This is done by defining “prefixes”
  - E.g. use “ctext:” as an abbreviation for https://data.ctext.org/entity/
    - Then to refer to “朱熹” (https://data.ctext.org/entity/597351) we instead just write “ctext:597351”



RDF data model: assigning globally unique identifiers

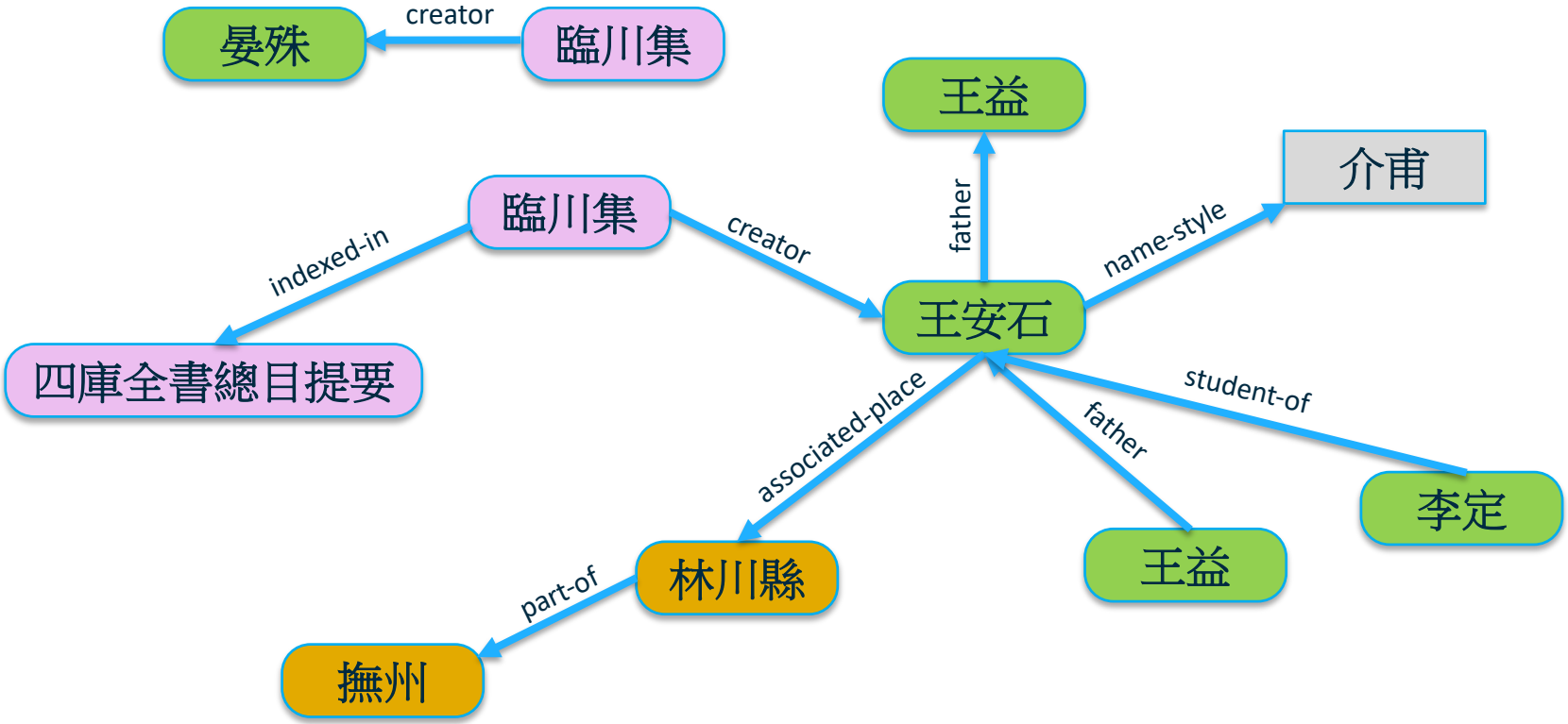
- To make things easier to read, we abbreviate the URIs
  - This is done by defining “prefixes”
  - This also makes it easy to refer to concepts defined by *others*

Defined by ctext

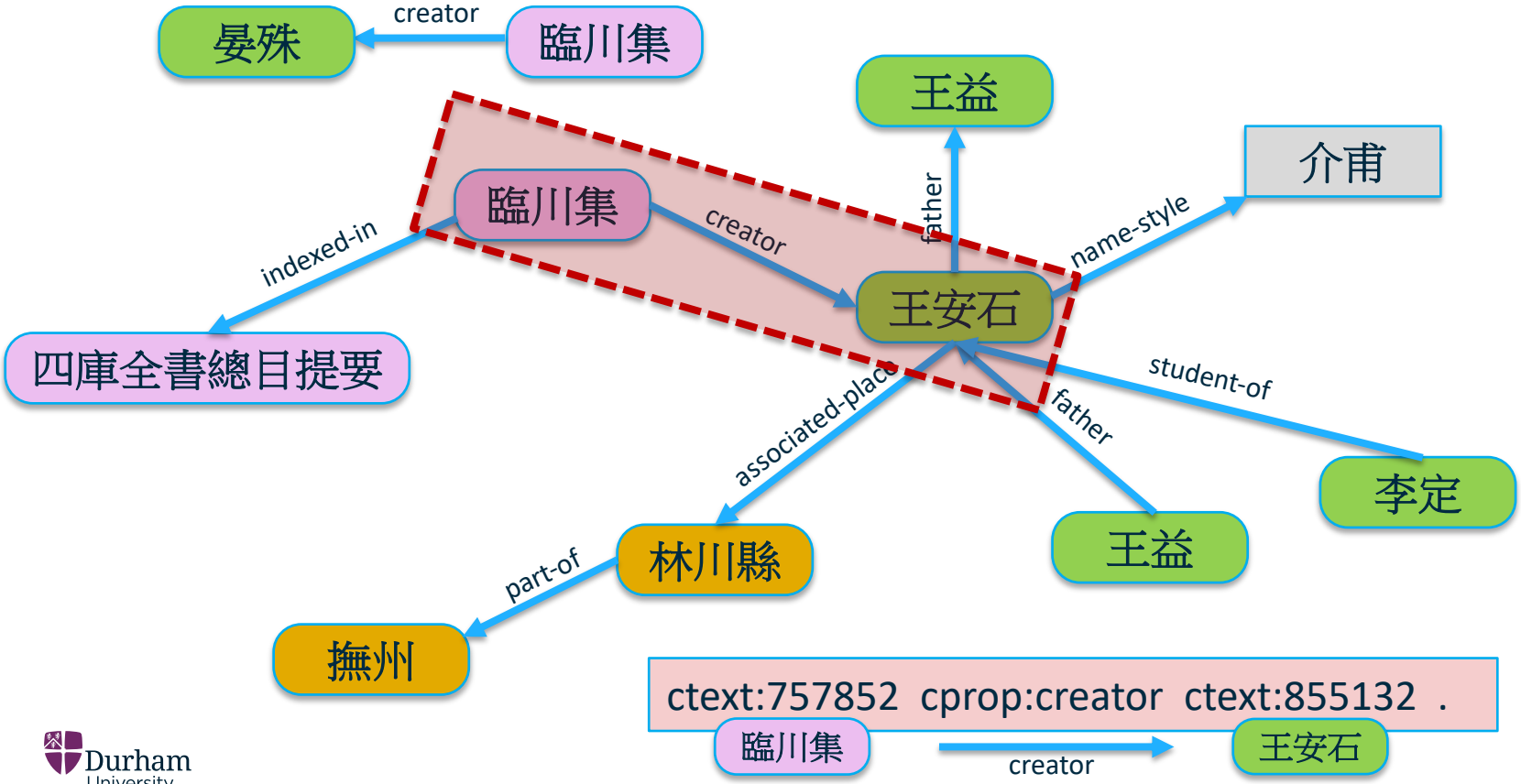
Defined by W3C

```
PREFIX date: <https://data.ctext.org/date/>
PREFIX cstat: <https://data.ctext.org/statement/>
PREFIX cqual: <https://data.ctext.org/qualifier/>
PREFIX cprop: <https://data.ctext.org/property/>
PREFIX claim: <https://data.ctext.org/claim/>
PREFIX ctext: <https://data.ctext.org/entity/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX time: <http://www.w3.org/2006/time#>
```

RDF data model: representing all data as a graph



RDF data model: representing all data as a graph

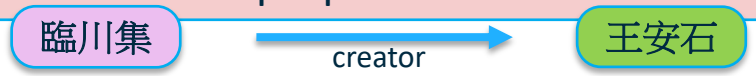


RDF data model and SPARQL

How data is expressed in RDF/SPARQL

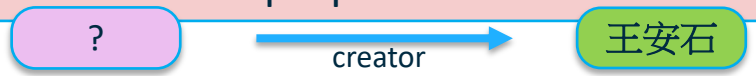
Pseudo-code explanation

ctext:757852 cprop:creator ctext:855132 .



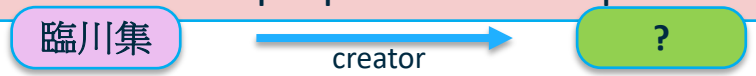
“臨川集 creator 王安石”

?work cprop:creator ctext:855132 .



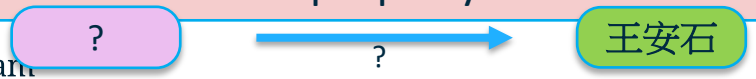
“??? creator 王安石”

ctext:757852 cprop:creator ?person .



“臨川集 creator ???”

?work ?property ctext:855132 .



“??? ??? 王安石”

E.g. “王雱 father 王安石”

Durham University

Advanced querying with SPARQL

SPARQL is the standard query language for RDF

Analogous to the SQL language for relational databases

(Further details/examples: SPARQL querying for ctext.org data)

Return *which* variables (* means "all those mentioned in the query")

```
SELECT * WHERE {
  ?subject ?verb ctext:813798 .
}
```

Variables (starting with "?") must match *these* statements

邵長蘅

ctext:813798

關係	對象
type	person
name	邵長蘅
name-style	子湘

Durham University

SPARQL, RDF, and prefixes

In the RDF model, every entity is identified by a unique URI

- E.g. https://data.ctext.org/entity/813798

In our query, we wrote “ctext:813798” instead – why does this work?

- We defined a *prefix* named “ctext:” that adds the rest of the URI!

```
1 ▸ PREFIX ↔
9 ▾ SELECT * WHERE {
10    ?subject ?verb ctext:813798 .
11 }
```

```
1 ▾ PREFIX date: <https://data.ctext.org/date/>
2   PREFIX cstat: <https://data.ctext.org/statement/>
3   PREFIX cqual: <https://data.ctext.org/qualifier/>
4   PREFIX cprop: <https://data.ctext.org/property/>
5   PREFIX claim: <https://data.ctext.org/claim/>
6   PREFIX ctext: <https://data.ctext.org/entity/>
7   PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
8   PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
9 ▾ SELECT * WHERE {
10    ?subject ?verb ctext:813798 .
11 }
```