Durham
University

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

Online materials for this session:
https://dsturgeon.net/hkadh2026

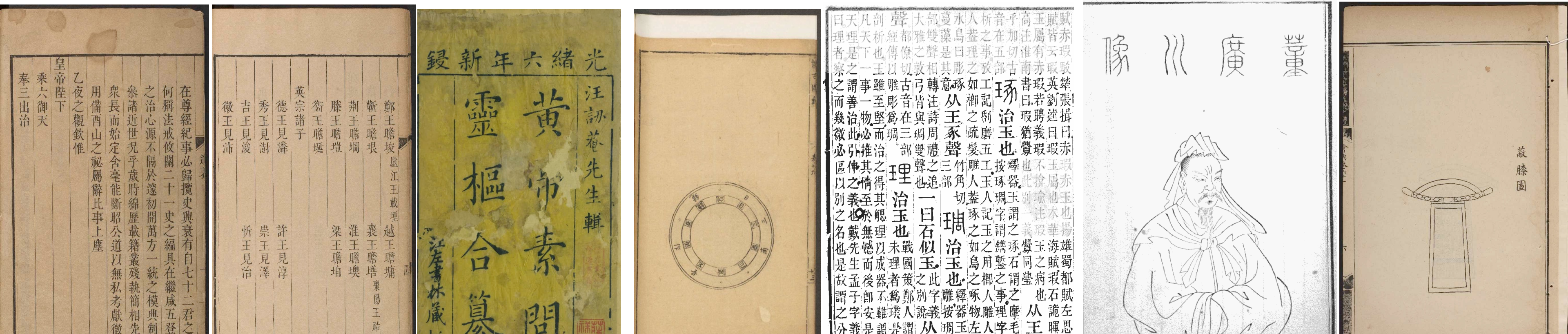
Donald Sturgeon
Department of Computer Science
Durham University, UK
donald.j.sturgeon@durham.ac.uk

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; regular expressions; interactive visualizations; …
- Historical data: knowledge graphs and semantic annotation
- RDF export and SPARQL search

Background: “Chinese Text Project” – ctext.org

- Online, open full-text database of premodern Chinese writing
- Online 2005-present
- Premodern transmitted editions of written works, Warring States – Qing
- Collectanea (叢書), works/editions published pre-20th century, etc.
    - 四庫全書、武英殿、四部叢刊、世界書局…



在尊經紀事必歸攬史興衰有自七十二君之
何稱法戒攸關二十一史之編具在繼咸五登
之治心源不隔於邃初開萬方一統之模典制
叅諸近世況乎歲時綿歷載籍叢殘執簡相先
衆長而始定含毫能斷昭公道以無私考獻徵
用備酉山之祕屬辭比事上塵
乙夜之觀欽惟
皇帝陛下
乘六御天
奉三出治

鄭王瞻埈（盧江王載堚）越王瞻墉
蘄王瞻垠　襄王瞻墡（東陽王祐
荆王瞻堈　淮王瞻墺
滕王瞻塏　梁王瞻垍
衛王瞻埏
英宗諸子
德王見潾　許王見淳
秀王見澍　崇王見澤
吉王見浚　忻王見治
徽王見沛

光緒六年新鐫
汪訒菴先生輯
黃帝素問
靈樞合纂
江左書林藏

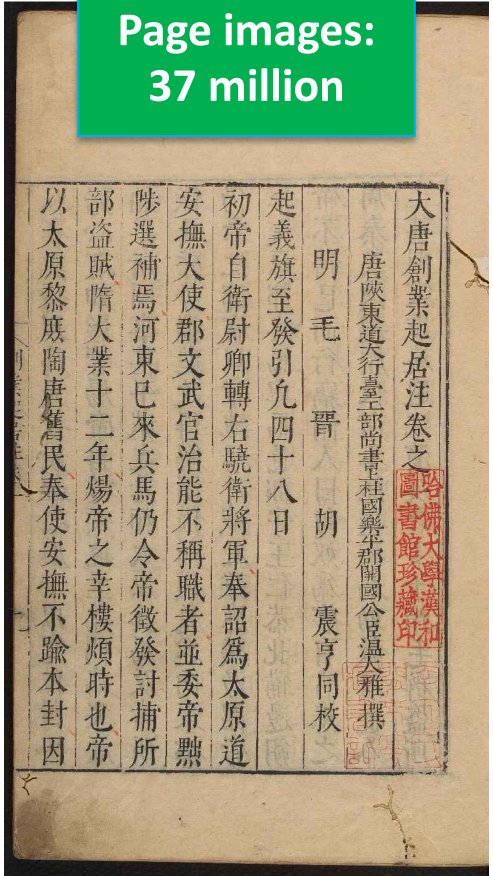
董廣川像

蔽膝圖

ctext.org – Most basic types of material

Primarily woodblock printed and handwritten sources in classical Chinese

Page images:
37 million



大唐創業起居注卷之一

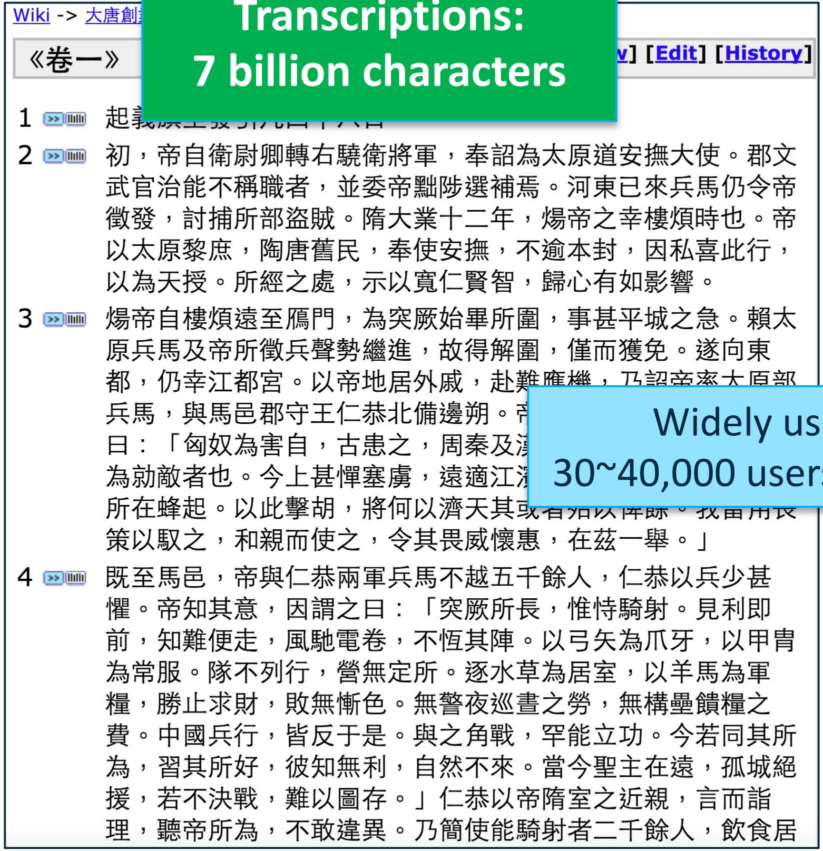
唐陝東道大行臺工部尚書上柱國樂平郡開國公臣溫大雅撰

明 毛晉 胡震亨 同校

起義旗至發引凡四十八日

初帝自衛尉卿轉右驍衛將軍奉詔爲太原道安撫大使郡文武官治能不稱職者並委帝黜陟選補焉河東已來兵馬仍令帝徵發討捕所部盜賊隋大業十二年煬帝之幸樓煩時也帝以太原黎庶陶唐舊民奉使安撫不踰本封因

哈佛大學漢和圖書館珍藏印



Wiki -> 大唐創

《卷一》

v] [Edit] [History]

1 起義

2 初，帝自衛尉卿轉右驍衛將軍，奉詔為太原道安撫大使。郡文武官治能不稱職者，並委帝黜陟選補焉。河東已來兵馬仍令帝徵發，討捕所部盜賊。隋大業十二年，煬帝之幸樓煩時也。帝以太原黎庶，陶唐舊民，奉使安撫，不逾本封，因私喜此行，以為天授。所經之處，示以寬仁賢智，歸心有如影響。

3 煬帝自樓煩遠至鴈門，為突厥始畢所圍，事甚平城之急。賴太原兵馬及帝所徵兵聲勢繼進，故得解圍，僅而獲免。遂向東都，仍幸江都宮。以帝地居外戚，赴難應機，乃詔帝率太原部兵馬，與馬邑郡守王仁恭北備邊朔。帝
日：「匈奴為害自，古患之，周秦及漢
為勍敵者也。今上甚憚塞虜，遠適江淮
所在蜂起。以此擊胡，將何以濟天其或者殆以俾歸。我當用長策以馭之，和親而使之，令其畏威懷惠，在茲一舉。」

4 既至馬邑，帝與仁恭兩軍兵馬不越五千餘人，仁恭以兵少甚懼。帝知其意，因謂之曰：「突厥所長，惟恃騎射。見利即前，知難便走，風馳電卷，不恒其陣。以弓矢為爪牙，以甲冑為常服。隊不列行，營無定所。逐水草為居室，以羊馬為軍糧，勝止求財，敗無慚色。無警夜巡晝之勞，無構壘饋糧之費。中國兵行，皆反于是。與之角戰，罕能立功。今若同其所為，習其所好，彼知無利，自然不來。當今聖主在遠，孤城絕援，若不決戰，難以圖存。」仁恭以帝隋室之近親，言而詣理，聽帝所為，不敢違異。乃簡使能騎射者二千餘人，飲食居

Transcriptions:
7 billion characters

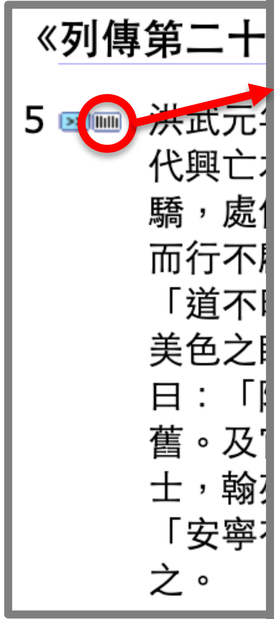
Widely used:
30~40,000 users per day

Durham University

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

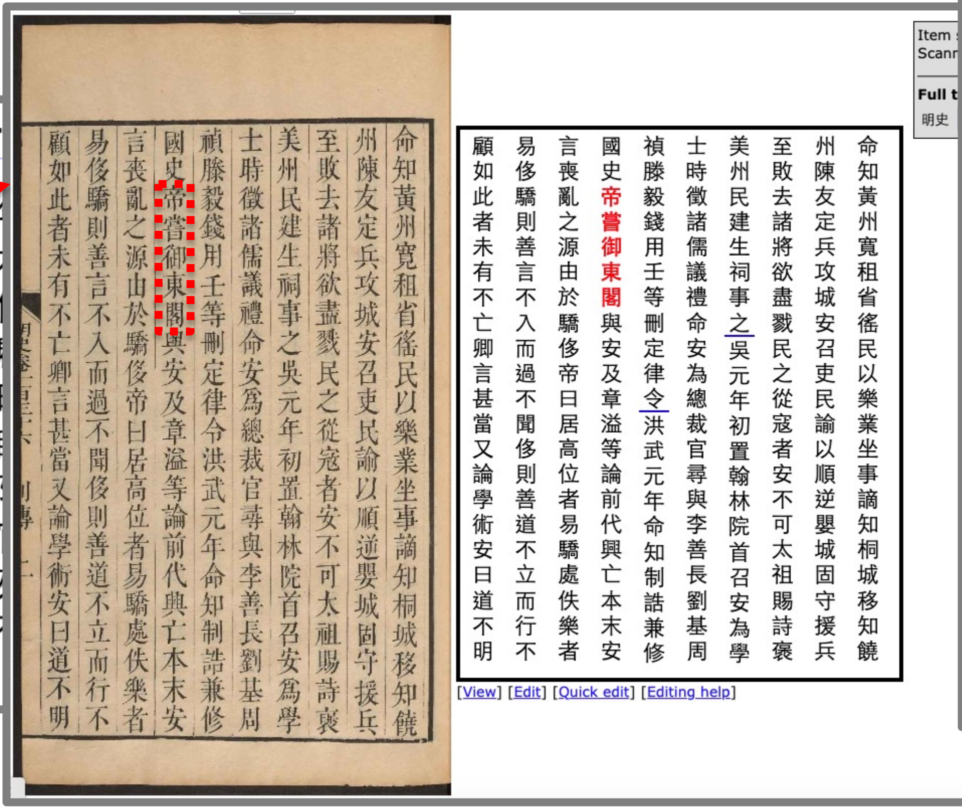
- A digital library and full-text database of premodern Chinese sources
- Main goals:
  - Organize premodern Chinese-language primary sources digitally
  - Create digital texts that are more than just surrogates for print
  - Do this in a way that scales to large volumes of material
    - Use automation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

Image sequences and literal transcriptions



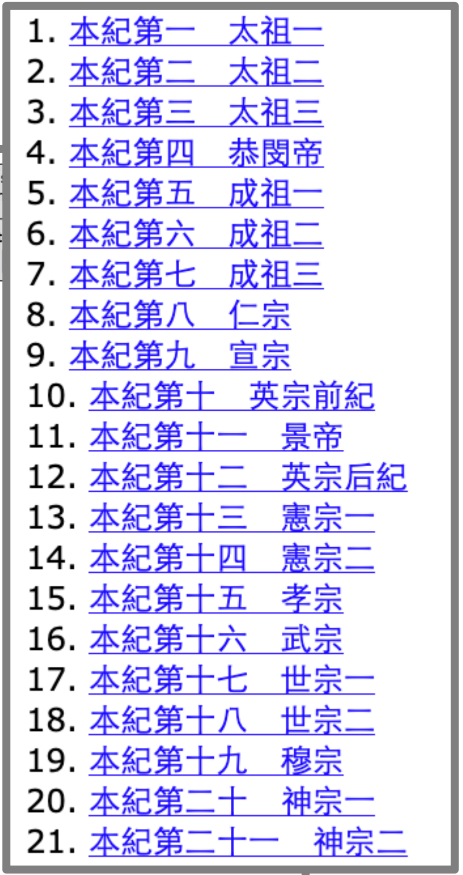
«列傳第二十

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

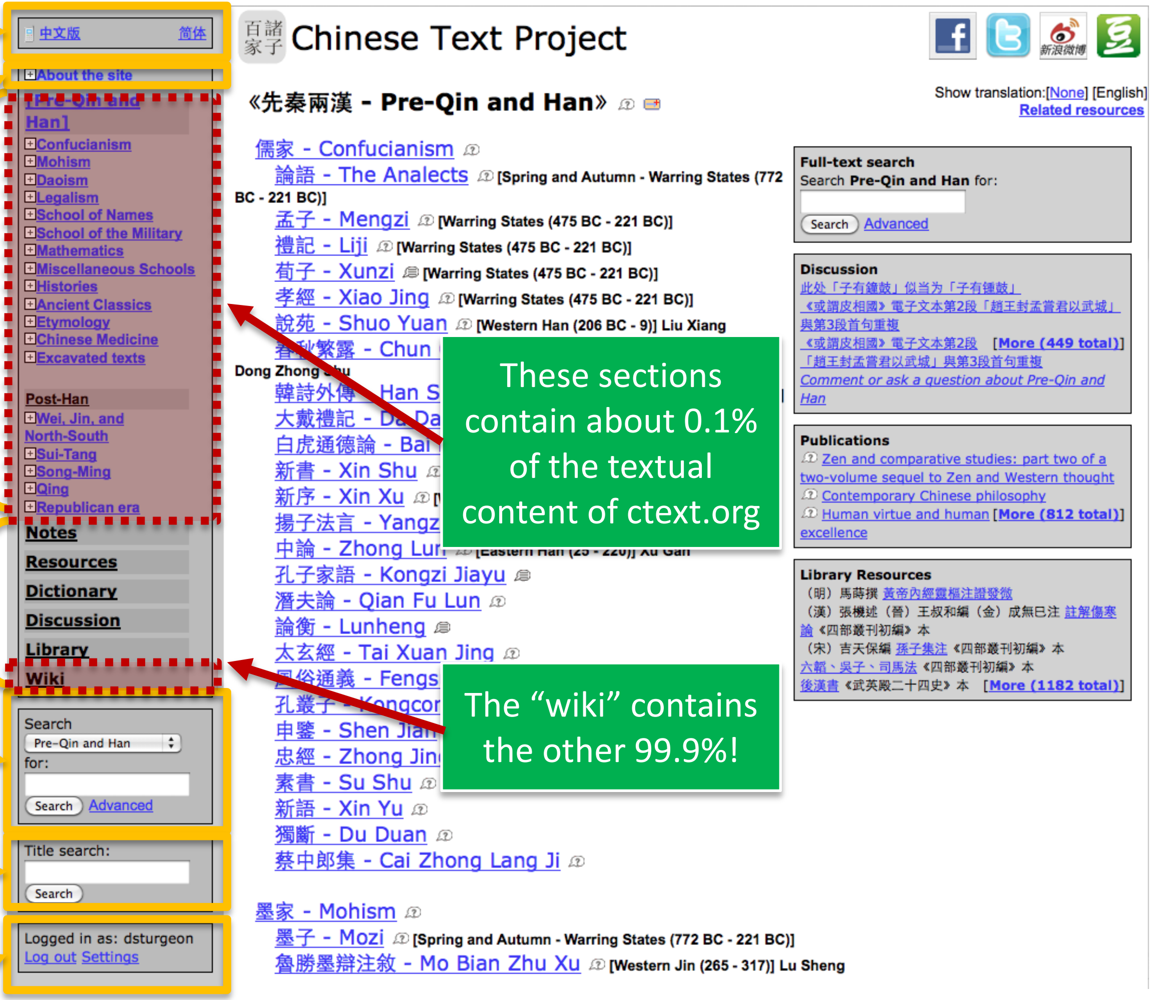


命知黃州寬租省傜民以樂業坐事謫知桐城移知饒
州陳友定兵攻城安召吏民諭以順逆嬰城固守援兵
至敗去諸將欲盡戮民之從寇者安不可太祖賜詩褒
美州民建生祠事之吳元年初置翰林院首召安為學
士時徵諸儒議禮命安為總裁官尋與李善長劉基周
禎滕毅錢用壬等刪定律令洪武元年命知制誥兼修
國史帝嘗御東閣與安及章溢等論前代興亡本末安
言喪亂之源由於驕侈帝日居高位者易驕處佚樂者
易侈驕則善言不入而過不聞侈則善道不立而行不
顧如此者未有不亡卿言甚當又論學術安日道不明

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



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



《先秦兩漢 - Pre-Qin and Han》

- 儒家 - Confucianism
  - 論語 - The Analects [Spring and Autumn - Warring States (772 BC - 221 BC)]
  - 孟子 - Mengzi [Warring States (475 BC - 221 BC)]
  - 禮記 - Liji [Warring States (475 BC - 221 BC)]
  - 荀子 - Xunzi [Warring States (475 BC - 221 BC)]
  - 孝經 - Xiao Jing [Warring States (475 BC - 221 BC)]
  - 說苑 - Shuo Yuan [Western Han (206 BC - 9)] Liu Xiang
  - 春秋繁露 - Chun Dong Zhong Shu
  - 韓詩外傳 - Han S
  - 大戴禮記 - Da Da
  - 白虎通德論 - Bai
  - 新書 - Xin Shu
  - 新序 - Xin Xu
  - 揚子法言 - Yangz
  - 中論 - Zhong Lun [Eastern Han (25 - 220)] Xu Gan
  - 孔子家語 - Kongzi Jiayu
  - 潛夫論 - Qian Fu Lun
  - 論衡 - Lunheng
  - 太玄經 - Tai Xuan Jing
  - 風俗通義 - Fengs
  - 孔叢子 - Kongcon
  - 申鑒 - Shen Jian
  - 忠經 - Zhong Jin
  - 素書 - Su Shu
  - 新語 - Xin Yu
  - 獨斷 - Du Duan
  - 蔡中郎集 - Cai Zhong Lang Ji
- 墨家 - Mohism
  - 墨子 - Mozi [Spring and Autumn - Warring States (772 BC - 221 BC)]
  - 魯勝墨辯注敘 - Mo Bian Zhu Xu [Western Jin (265 - 317)] Lu Sheng

Full-text search
Search **Pre-Qin and Han** for:

Discussion
此处「子有鐘鼓」似当为「子有锺鼓」
《或謂皮相國》電子文本第2段「趙王封孟嘗君以武城」與第3段首句重複
《或謂皮相國》電子文本第2段「趙王封孟嘗君以武城」與第3段首句重複 [**More (449 total)**]
Comment or ask a question about Pre-Qin and Han

Publications
Zen and comparative studies: part two of a two-volume sequel to Zen and Western thought
Contemporary Chinese philosophy
Human virtue and human excellence [**More (812 total)**]

Library Resources
（明）馬蒔撰 黃帝內經靈樞注證發微
（漢）張機述（晉）王叔和編（金）成無已注 註解傷寒論《四部叢刊初編》本
（宋）吉天保編 孫子集注《四部叢刊初編》本
六韜、吳子、司馬法《四部叢刊初編》本
後漢書《武英殿二十四史》本 [**More (1182 total)**]

These sections contain about 0.1% of the textual content of ctext.org

The “wiki” contains the other 99.9%!

Interface
中／英 繁／簡

Instructions

Textual database (pre-Qin and Han texts)

Other sections: Library, Wiki, Dictionary, etc.

Full-text search

Title search

Login & Settings

Durham University

Please take a note of these pieces of information:

1. API key:

2. Links, notes, and step-by-step tutorials:
   https://dsturgeon.net/hkadh2026

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:

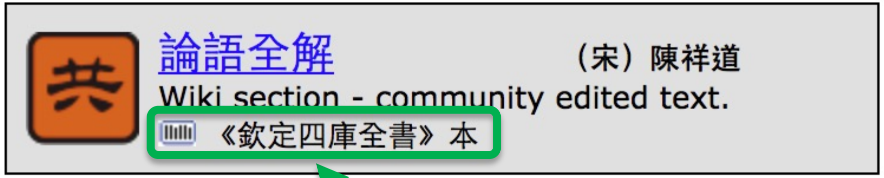
文 Transcription (text DB)
(not user editable)

共 Transcription (wiki)
(user editable)

共 Transcription (OCR, wiki)
(uncorrected, editable)

像 Scanned primary source
(not a transcription)

Example:



共 論語全解 （宋）陳祥道
Wiki section - community edited text.
《欽定四庫全書》本

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

Durham
University

Editing content

Maintained using a publicly editable, versioned system (i.e. a “wiki”)

Currently editable contents:

Type of content	Section of ctext.org	Similar to
Transcriptions of texts	=> “Wiki”	≈ Wikisource
Translations of texts	=> “Wiki”	
Data about historical entities	=> “Data Wiki”	≈ Wikidata
Descriptions of historical entities	=> “Text Wiki”	≈ Wikipedia

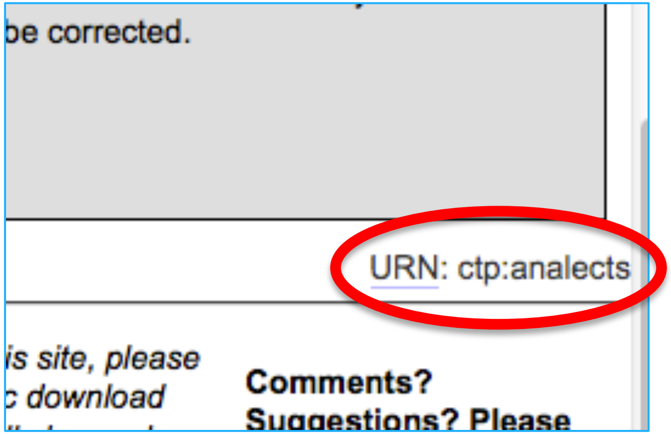
Hands-on tutorial: Part II
Text mining with ctext’s Text Tools plugin

- 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:…”

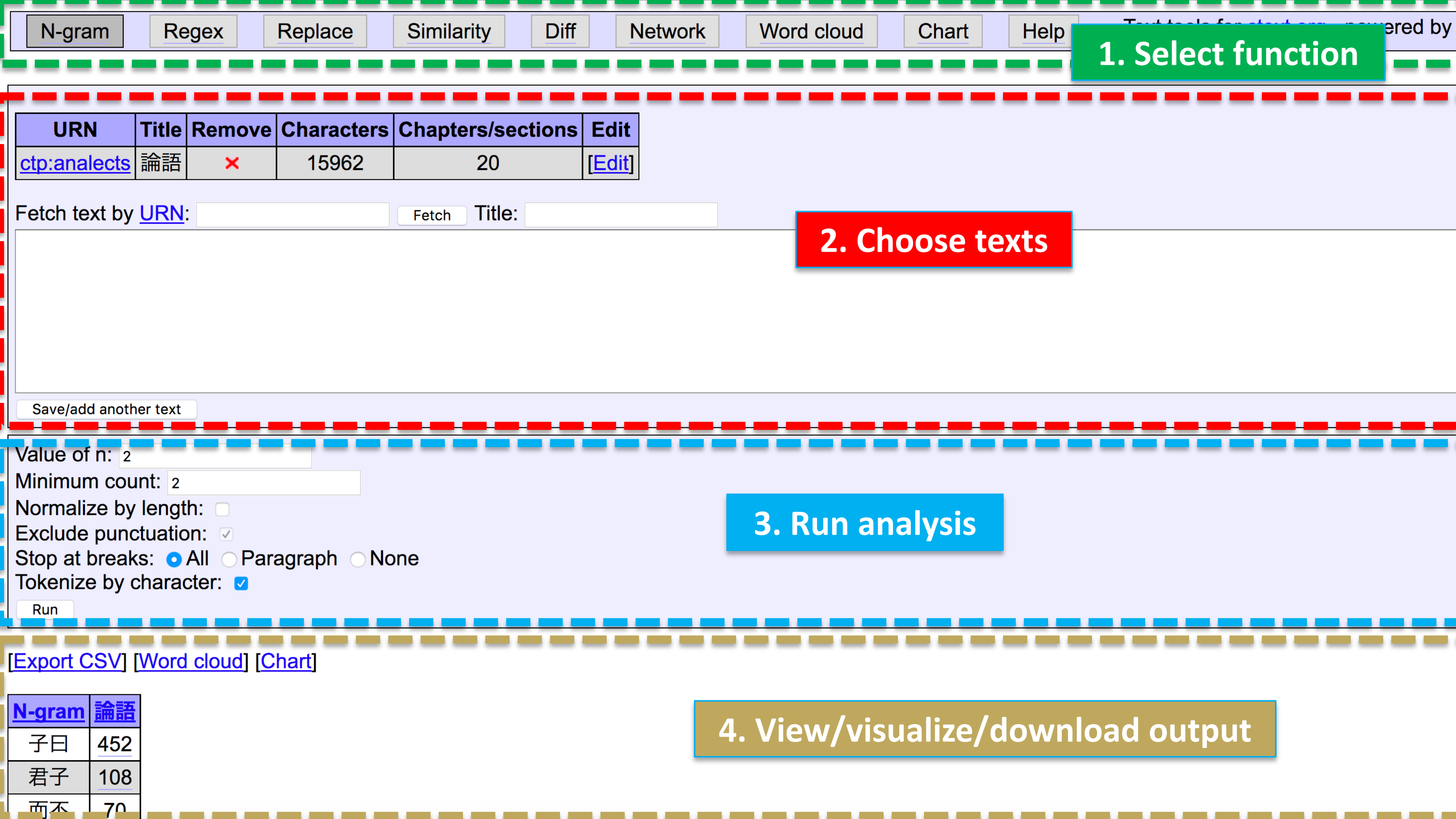
- To “decode”: same as finding texts by title:
  - Paste URN into “Title search box”
  - Click “Search”
  - Contents page for that text will open


be corrected.

URN: ctp:analects

is site, please
download

Comments?
Suggestions? Please



1. Select function

2. Choose texts

URN	Title	Remove	Characters	Chapters/sections	Edit
ctp:analects	論語	×	15962	20	[Edit]

Fetch text by URN:
Title:

3. Run analysis

Value of n: 2
Minimum count: 2
Normalize by length: ☐
Exclude punctuation: ☑
Stop at breaks: ● All ○ Paragraph ○ None
Tokenize by character: ☑

4. View/visualize/download output

N-gram	論語
子曰	452
君子	108
而不	70

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

Durham
University

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。

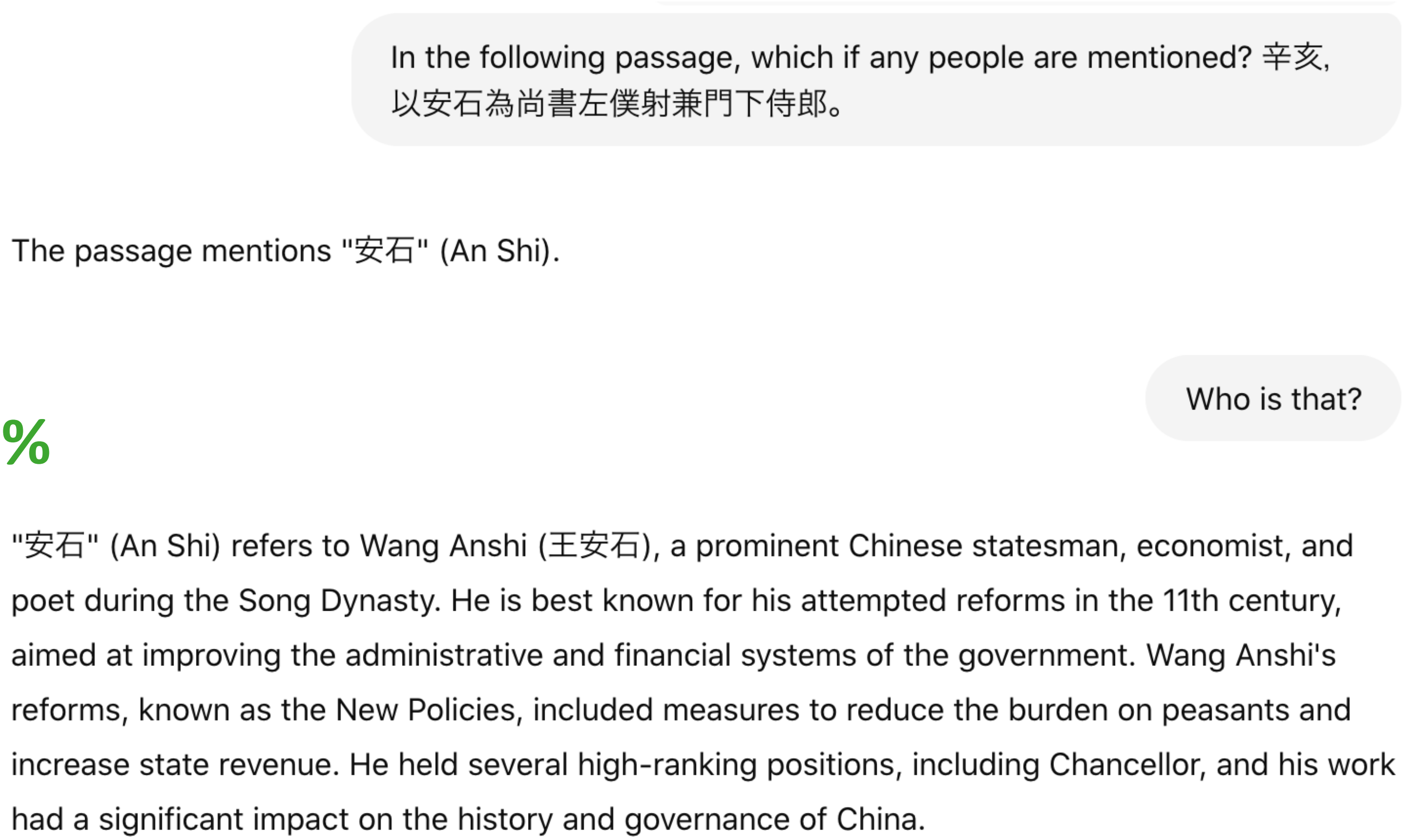
Later [he] died of illness. [He] authored *Bing Lue*, which was widely praised. [His] son was Anshi.

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

Can’t we just ChatGPT this?



In the following passage, which if any people are mentioned? 辛亥,以安石為尚書左僕射兼門下侍郎。

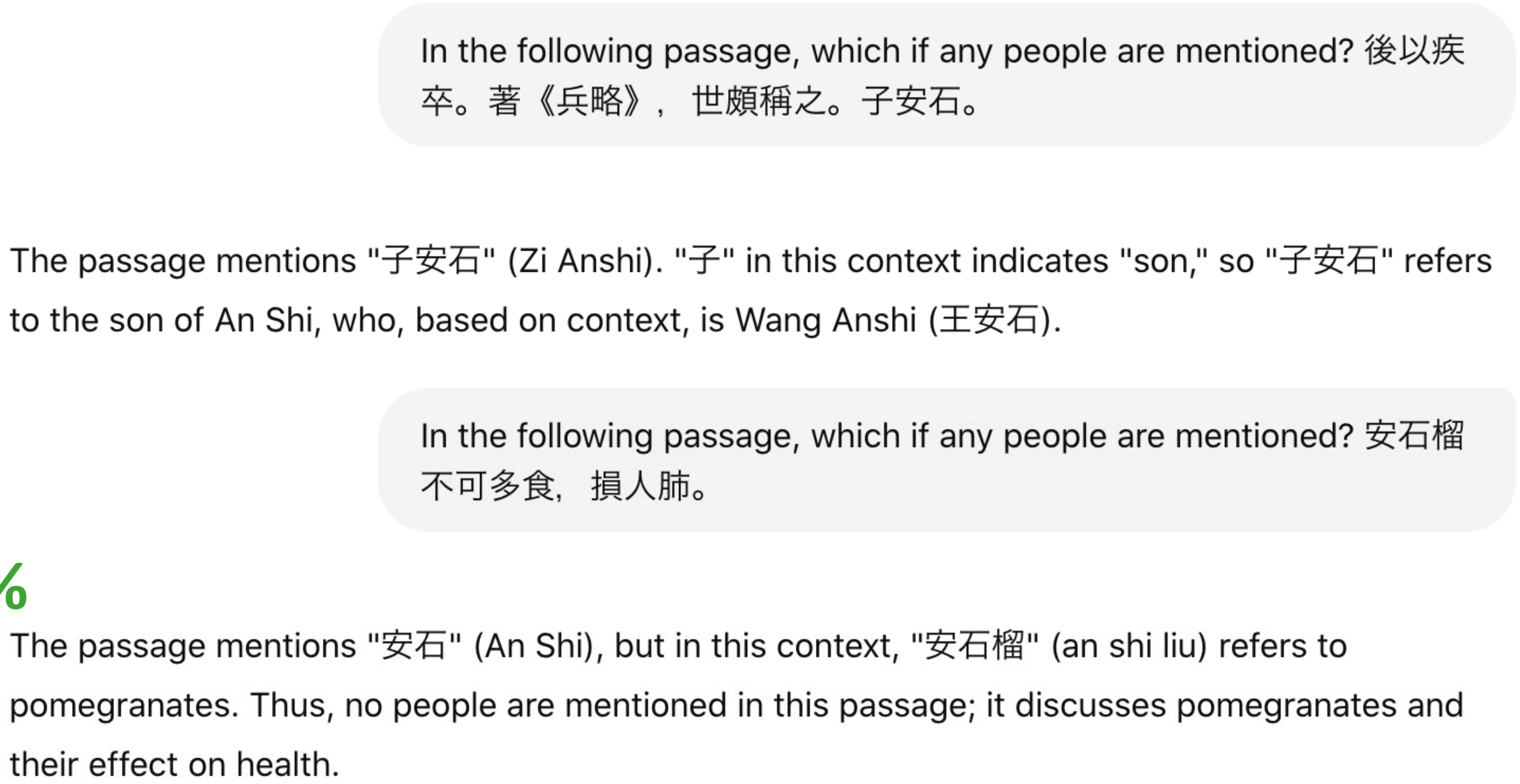
The passage mentions "安石" (An Shi).

Who is that?

✓100%

"安石" (An Shi) refers to Wang Anshi (王安石), a prominent Chinese statesman, economist, and poet during the Song Dynasty. He is best known for his attempted reforms in the 11th century, aimed at improving the administrative and financial systems of the government. Wang Anshi's reforms, known as the New Policies, included measures to reduce the burden on peasants and increase state revenue. He held several high-ranking positions, including Chancellor, and his work had a significant impact on the history and governance of China.

Can’t we just ChatGPT this?



In the following passage, which if any people are mentioned? 後以疾卒。著《兵略》，世頗稱之。子安石。

50%

The passage mentions "子安石" (Zi Anshi). "子" in this context indicates "son," so "子安石" refers to the son of An Shi, who, based on context, is Wang Anshi (王安石).

In the following passage, which if any people are mentioned? 安石榴不可多食，損人肺。

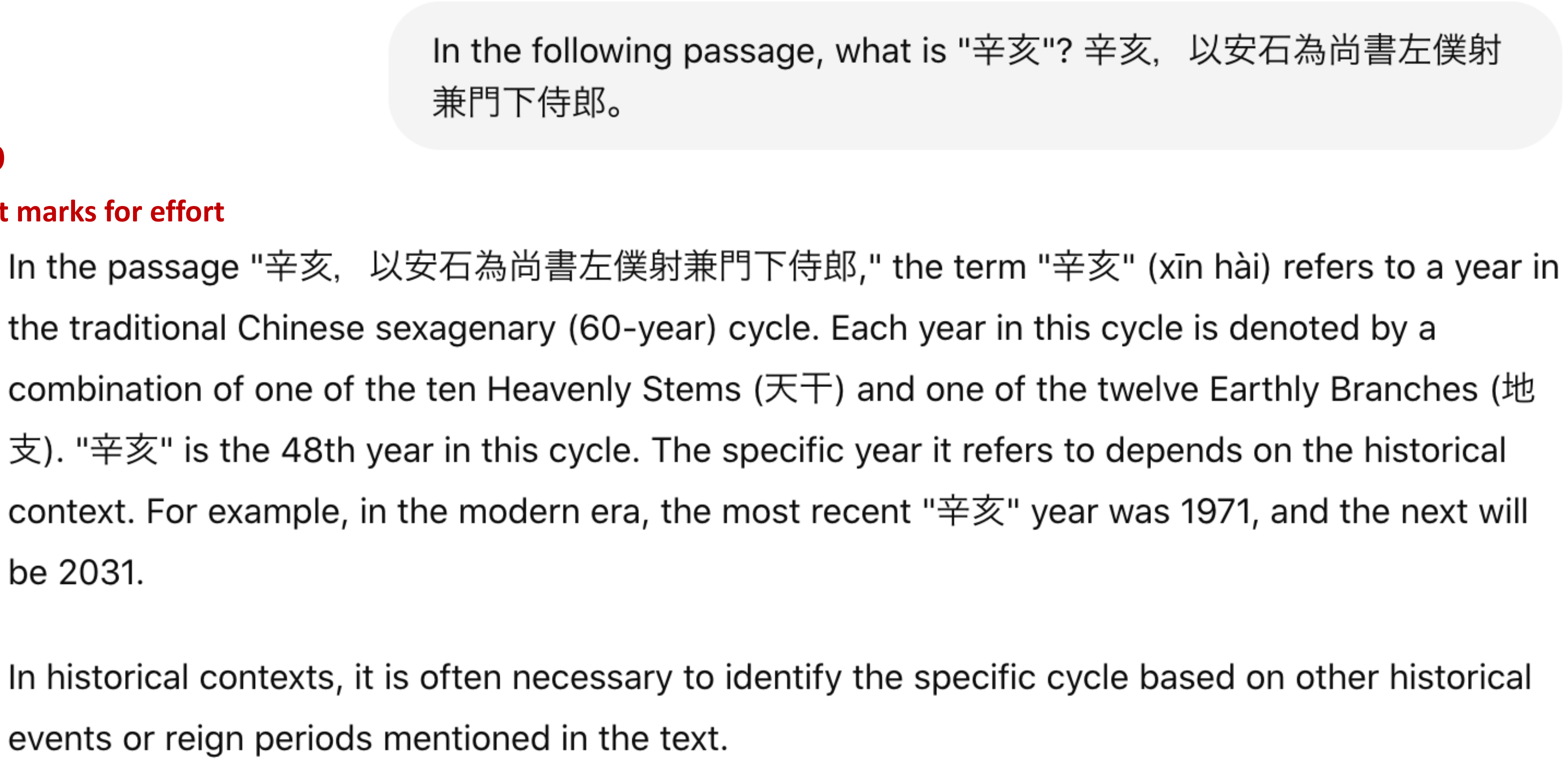
✓100%

The passage mentions "安石" (An Shi), but in this context, "安石榴" (an shi liu) refers to pomegranates. Thus, no people are mentioned in this passage; it discusses pomegranates and their effect on health.

Can’t we just ChatGPT this?

10%

Wrong but marks for effort

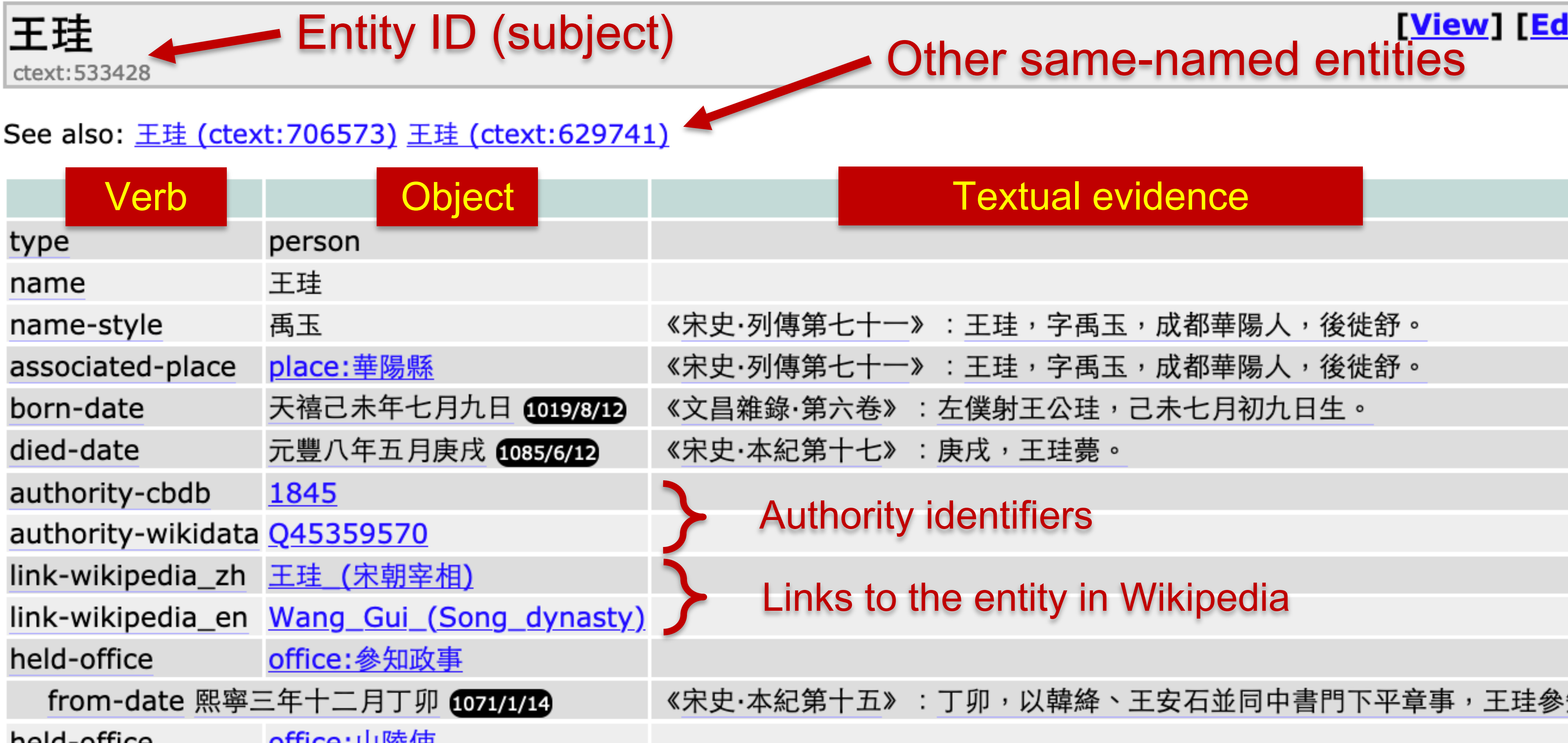


In the following passage, what is "辛亥"? 辛亥，以安石為尚書左僕射兼門下侍郎。

In the passage "辛亥，以安石為尚書左僕射兼門下侍郎," the term "辛亥" (xīn hài) refers to a year in the traditional Chinese sexagenary (60-year) cycle. Each year in this cycle is denoted by a combination of one of the ten Heavenly Stems (天干) and one of the twelve Earthly Branches (地支). "辛亥" is the 48th year in this cycle. The specific year it refers to depends on the historical context. For example, in the modern era, the most recent "辛亥" year was 1971, and the next will be 2031.

In historical contexts, it is often necessary to identify the specific cycle based on other historical events or reign periods mentioned in the text.

Entity records and knowledge claims



王珪
ctext:533428

[View] [Ed

Entity ID (subject)

Other same-named entities

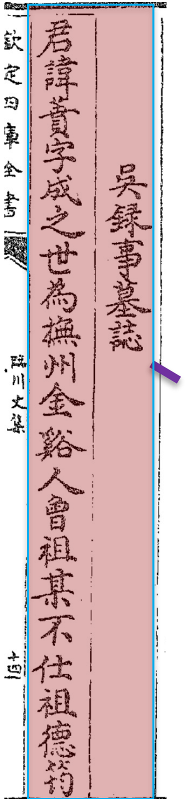
See also: 王珪 (ctext:706573) 王珪 (ctext:629741)

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-wikidata	Q45359570	
link-wikipedia_zh	王珪_(宋朝宰相)	
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:山陵使	

Authority identifiers

Links to the entity in Wikipedia

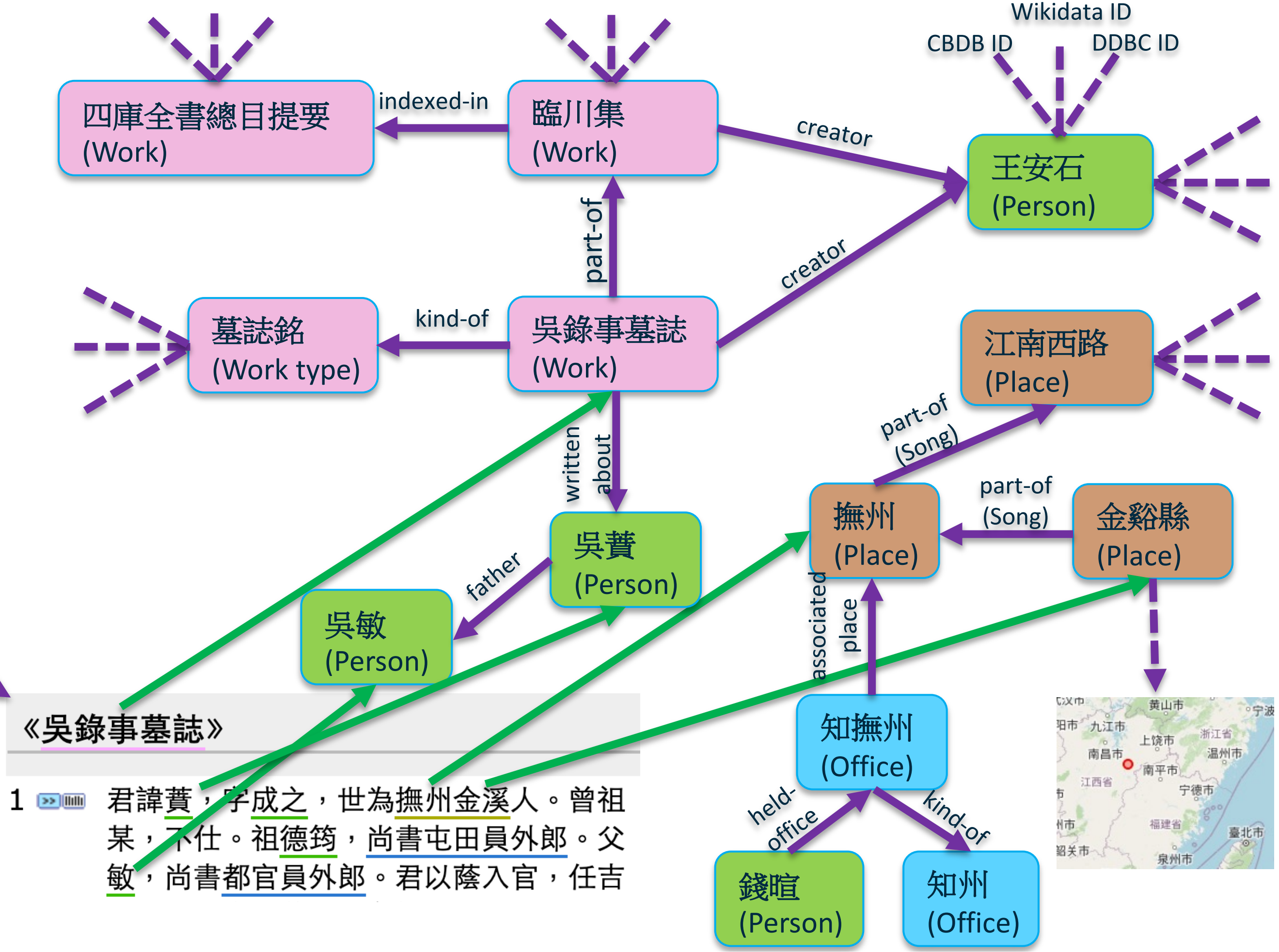
Semantic annotation & knowledge graph



吳錄事墓誌

君諱蕢字成之世為撫州金谿人曾祖某不仕祖德筠

欽定四庫全書
臨川文集
十四



《吳錄事墓誌》

1 君諱蕢，字成之，世為撫州金溪人。曾祖某，不仕。祖德筠，尚書屯田員外郎。父敏，尚書都官員外郎。君以蔭入官，任吉

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

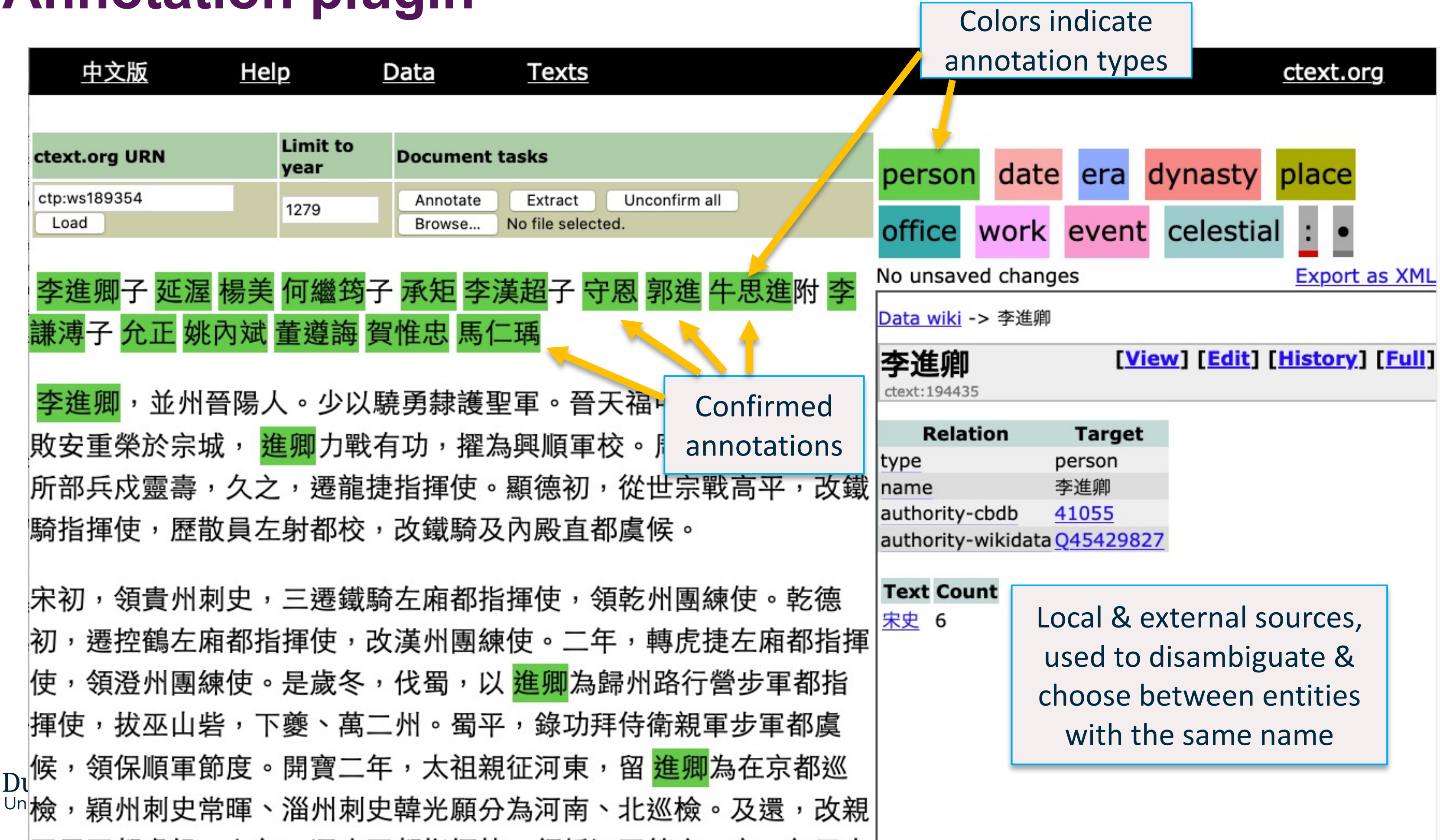
Data: basic searching

- 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



Colors indicate annotation types

ctext.org URN	Limit to year	Document tasks
ctp:ws189354	1279	No file selected.

李進卿子 延渥 楊美 何繼筠子 承矩 李漢超子 守恩 郭進 牛思進附 李謙溥子 允正 姚內斌 董遵誨 賀惟忠 馬仁瑀

李進卿，並州晉陽人。少以驍勇隸護聖軍。晉天福

敗安重榮於宗城，進卿力戰有功，擢為興順軍校。

所部兵戍靈壽，久之，遷龍捷指揮使。顯德初，從世宗戰高平，改鐵騎指揮使，歷散員左射都校，改鐵騎及內殿直都虞候。

宋初，領貴州刺史，三遷鐵騎左廂都指揮使，領乾州團練使。乾德初，遷控鶴左廂都指揮使，改漢州團練使。二年，轉虎捷左廂都指揮使，領澄州團練使。是歲冬，伐蜀，以進卿為歸州路行營步軍都指揮使，拔巫山砦，下夔、萬二州。蜀平，錄功拜侍衛親軍步軍都虞候，領保順軍節度。開寶二年，太祖親征河東，留進卿為在京都巡檢，潁州刺史常暉、淄州刺史韓光願分為河南、北巡檢。及還，改親

Confirmed annotations

No unsaved changes

Data wiki -> 李進卿

李進卿

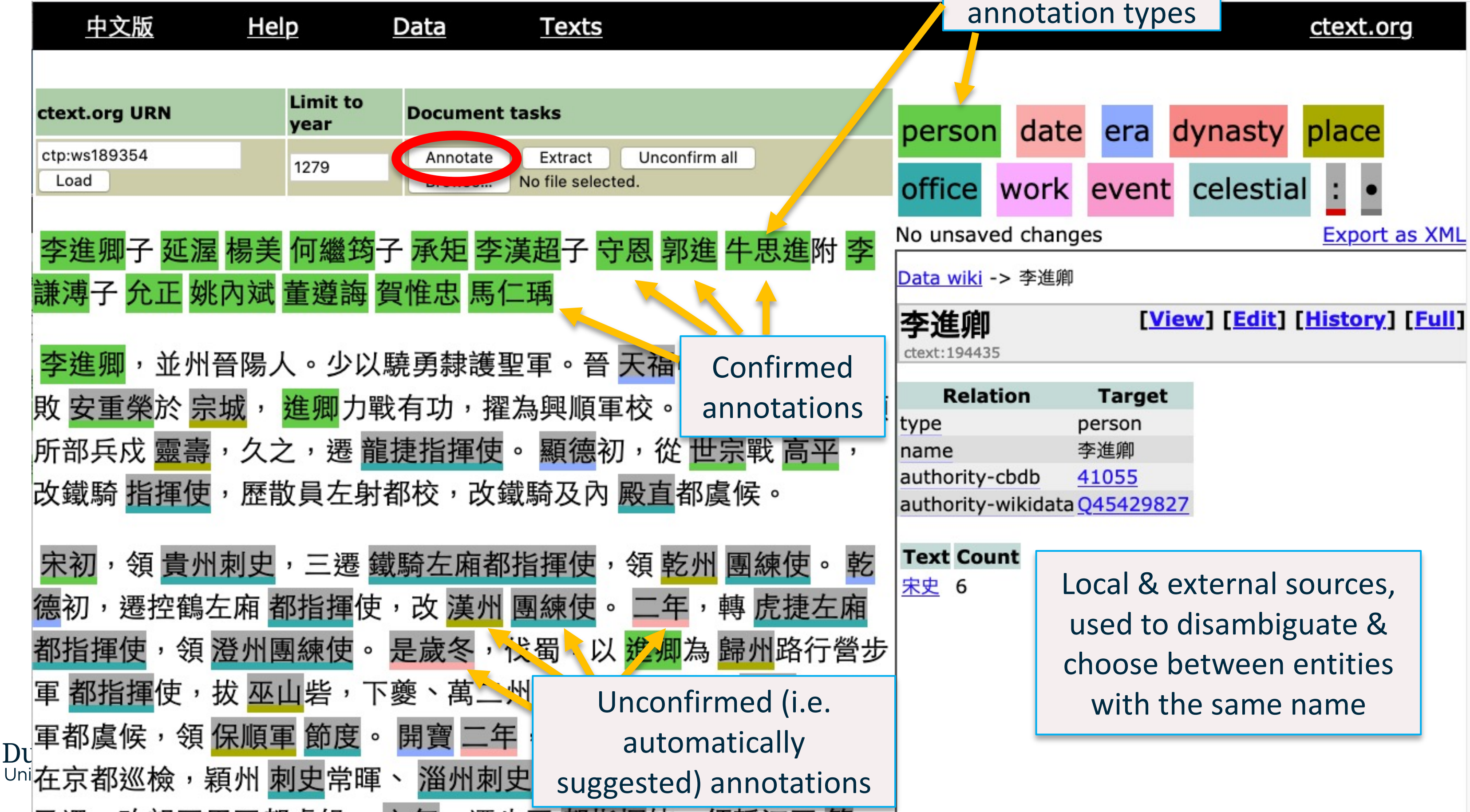
ctext:194435

Relation	Target
type	person
name	李進卿
authority-cbdb	41055
authority-wikidata	Q45429827

Text	Count
宋史	6

Local & external sources, used to disambiguate & choose between entities with the same name

Annotation plugin



ctext.org URN	Limit to year	Document tasks
ctp:ws189354 Load	1279	Annotate Extract Unconfirm all Browse... No file selected.

No unsaved changes

Export as XML

Data wiki -> 李進卿

李進卿 [View] [Edit] [History] [Full]

ctext:194435

Relation	Target
type	person
name	李進卿
authority-cbdb	41055
authority-wikidata	Q45429827

Text	Count
宋史	6

李進卿子 延渥 楊美 何繼筠子 承矩 李漢超子 守恩 郭進 牛思進附 李謙溥子 允正 姚內斌 董遵誨 賀惟忠 馬仁瑀

李進卿，並州晉陽人。少以驍勇隸護聖軍。晉 天福
敗 安重榮於 宗城， 進卿力戰有功，擢為興順軍校。
所部兵戍 靈壽，久之，遷 龍捷指揮使。 顯德初，從 世宗戰 高平，
改鐵騎 指揮使，歷散員左射都校，改鐵騎及內 殿直都虞候。

宋初，領 貴州刺史，三遷 鐵騎左廂都指揮使，領 乾州 團練使。 乾德初，遷控鶴左廂 都指揮使，改 漢州 團練使。 二年，轉 虎捷左廂
都指揮使，領 澄州團練使。 是歲冬，伐蜀，以 進卿為 歸州路行營步
軍 都指揮使，拔 巫山砦，下夔、萬
軍都虞候，領 保順軍 節度。 開寶 二年
在京都巡檢，穎州 刺史常暉、 淄州刺史

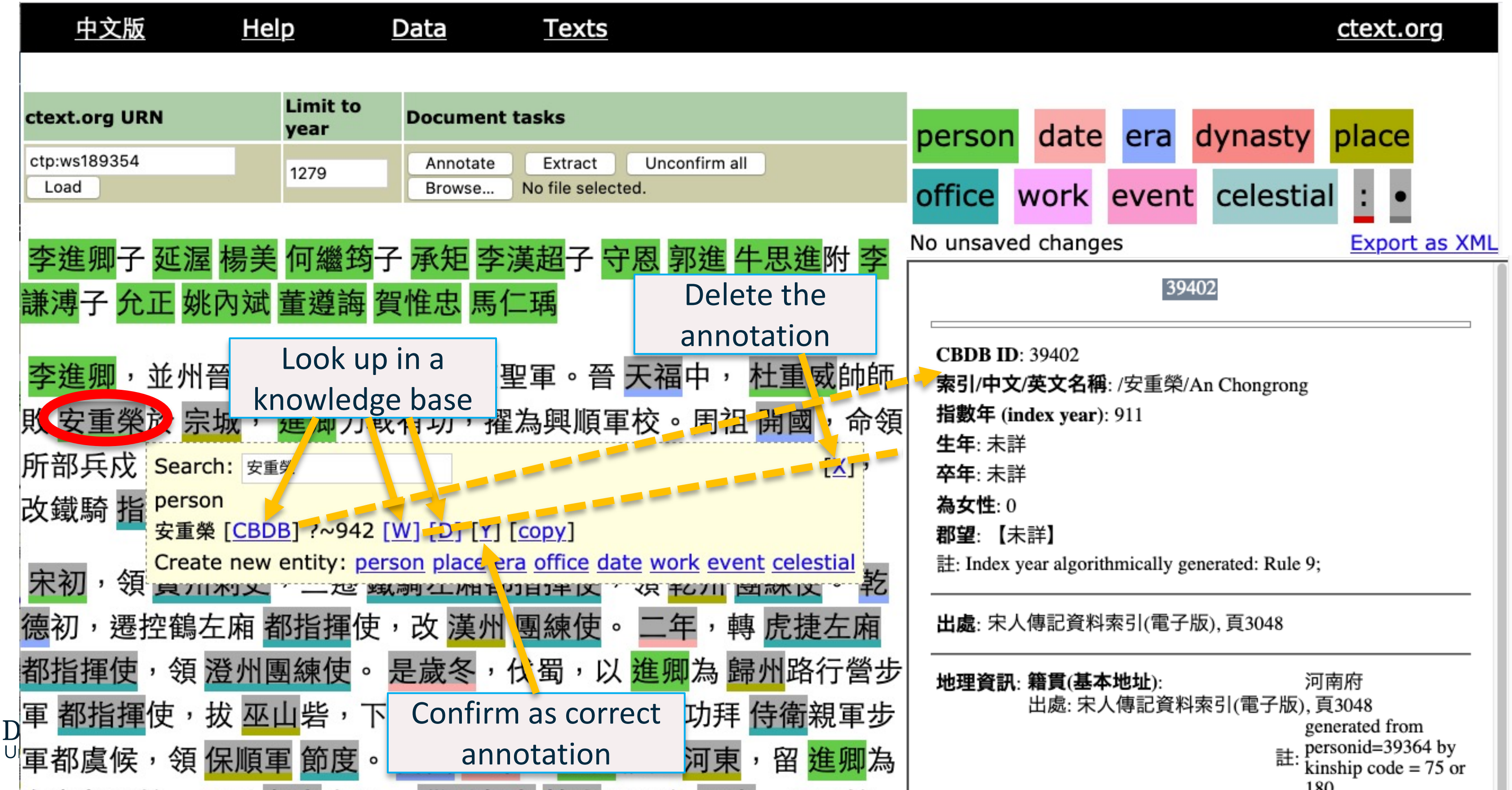
Colors indicate annotation types

Confirmed annotations

Unconfirmed (i.e. automatically suggested) annotations

Local & external sources, used to disambiguate & choose between entities with the same name

Annotation plugin



No unsaved changes

李進卿子 延渥 楊美 何繼筠子 承矩 李漢超子 守恩 郭進 牛思進附 李謙溥子 允正 姚內斌 董遵誨 賀惟忠 馬仁瑀

李進卿，並州晉 聖軍。晉 天福中， 杜重威帥師敗 安重榮於 宗城， 進卿力戰有功，擢為興順軍校。周祖 開國，命領所部兵戍

改鐵騎 指

宋初，領 乾德初，遷控鶴左廂 都指揮使，改 漢州 團練使。 二年，轉 虎捷左廂都指揮使，領 澄州團練使。 是歲冬，伐蜀，以 進卿為 歸州路行營步軍 都指揮使，拔 巫山砦，下 功拜 侍衛親軍步軍都虞候，領 保順軍 節度。 河東，留 進卿為

Search: 安重榮 [X]

person

安重榮 [CBDB] ?~942 [W] [D] [Y] [copy]

Create new entity: person place era office date work event celestial

39402

CBDB ID: 39402

索引/中文/英文名稱: /安重榮/An Chongrong

指數年 (index year): 911

生年: 未詳

卒年: 未詳

為女性: 0

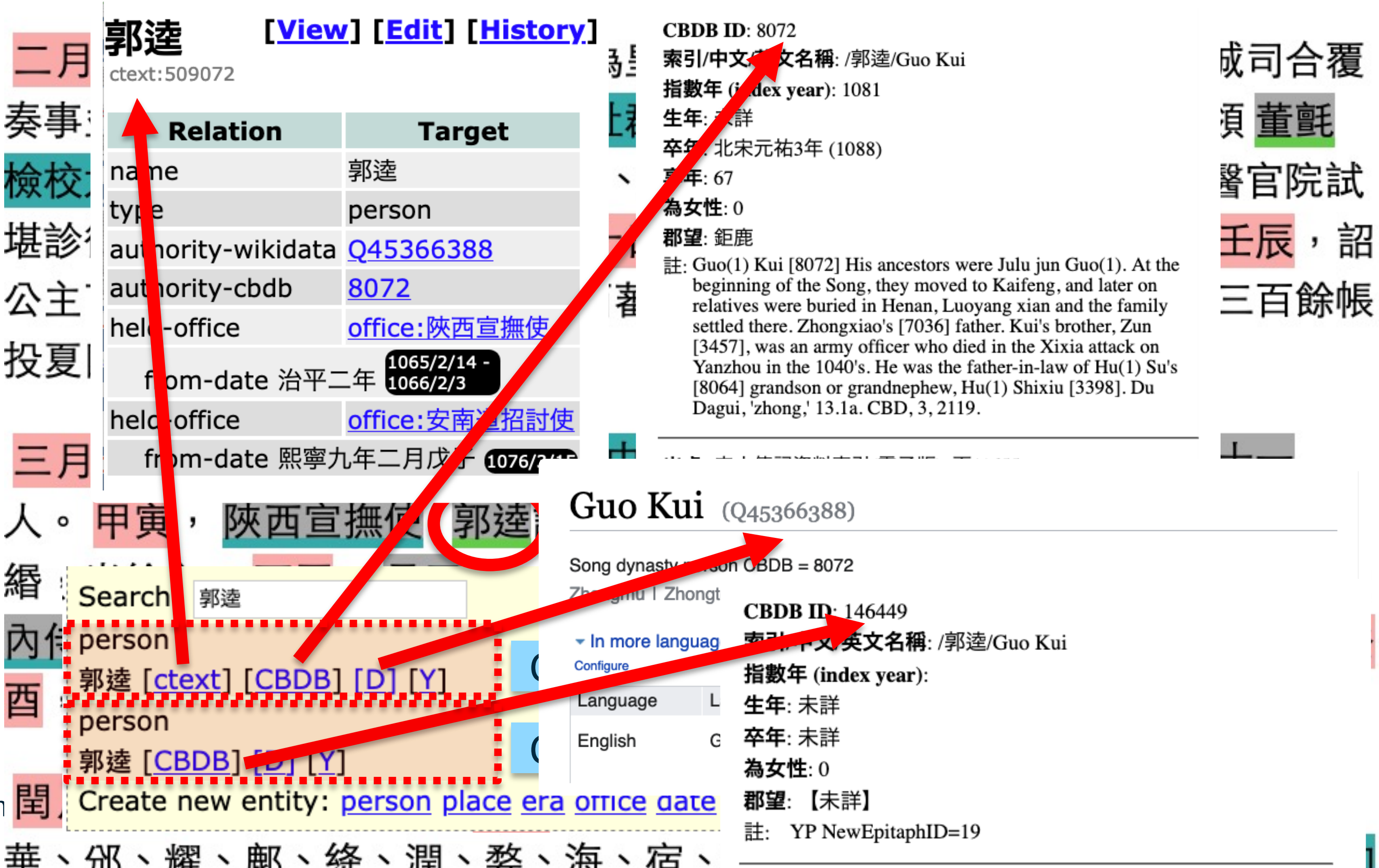
郡望: 【未詳】

註: Index year algorithmically generated: Rule 9;

出處: 宋人傳記資料索引(電子版), 頁3048

地理資訊: **籍貫(基本地址)**: 河南府
出處: 宋人傳記資料索引(電子版), 頁3048
註: generated from personid=39364 by kinship code = 75 or 180

Annotation plugin



郭逵 [View] [Edit] [History]

ctext:509072

Relation	Target
name	郭逵
type	person
authority-wikidata	Q45366388
authority-cbdb	8072
held-office	office:陝西宣撫使
from-date	治平二年 1065/2/14 - 1066/2/3
held-office	office:安南道招討使
from-date	熙寧九年二月戊午 1076/2/15

人。甲寅，陝西宣撫使郭逵

Search 郭逵

person
郭逵 [ctext] [CBDB] [D] [Y]

person
郭逵 [CBDB] [D] [Y]

Create new entity: person place era office date

CBDB ID: 8072
索引/中文/英文名稱: /郭逵/Guo Kui
指數年 (index year): 1081
生年: 未詳
卒年: 北宋元祐3年 (1088)
享年: 67
為女性: 0
郡望: 鉅鹿
註: Guo(1) Kui [8072] His ancestors were Julu jun Guo(1). At the beginning of the Song, they moved to Kaifeng, and later on relatives were buried in Henan, Luoyang xian and the family settled there. Zhongxiao's [7036] father. Kui's brother, Zun [3457], was an army officer who died in the Xixia attack on Yanzhou in the 1040's. He was the father-in-law of Hu(1) Su's [8064] grandson or grandnephew, Hu(1) Shixiu [3398]. Du Dagui, 'zhong,' 13.1a. CBD, 3, 2119.

Guo Kui (Q45366388)

Song dynasty person CBDB = 8072

CBDB ID: 146449
索引/中文/英文名稱: /郭逵/Guo Kui
指數年 (index year):
生年: 未詳
卒年: 未詳
為女性: 0
郡望: 【未詳】
註: YP NewEpitaphID=19

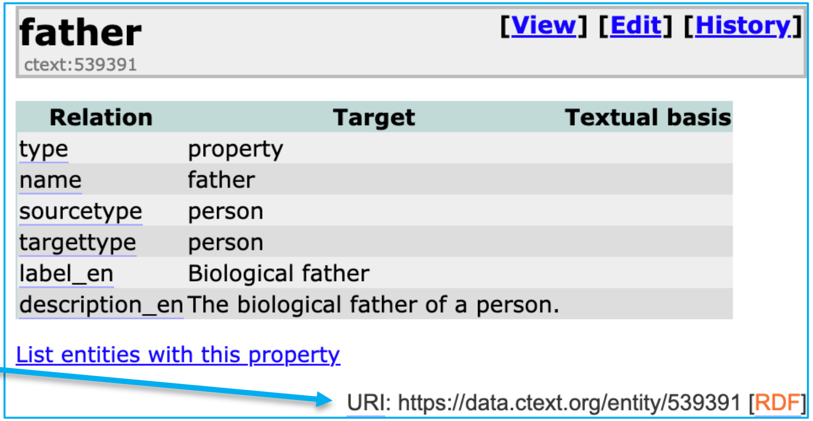
Durham University

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:



father [View] [Edit] [History]

ctext:539391

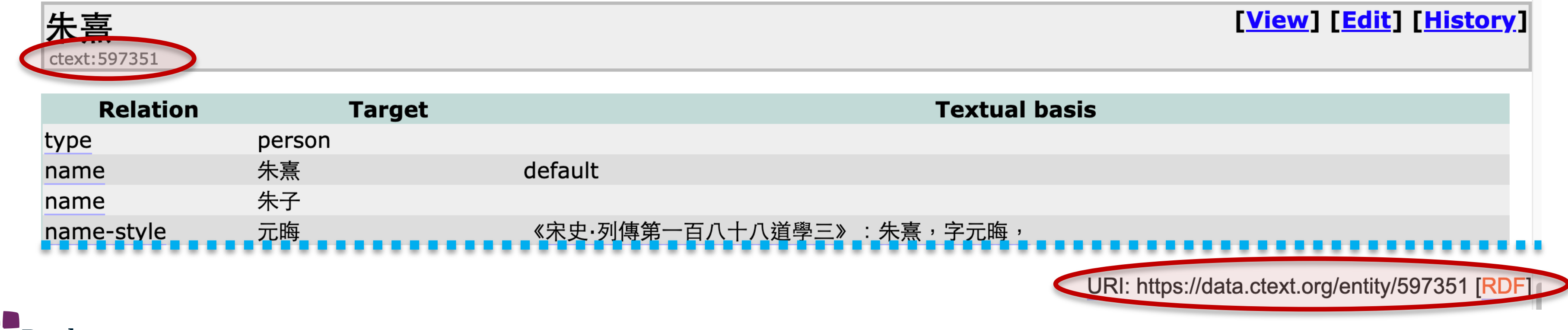
Relation	Target	Textual basis
type	property	
name	father	
sourcetype	person	
targettype	person	
label_en	Biological father	
description_en	The biological father of a person.	

List entities with this property

URI: https://data.ctext.org/entity/539391 [RDF]

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”



朱熹

ctext:597351

[View] [Edit] [History]

Relation	Target	Textual basis
type	person	
name	朱熹	default
name	朱子	
name-style	元晦	《宋史·列傳第一百八十八道學三》：朱熹，字元晦，

URI: https://data.ctext.org/entity/597351 [RDF]

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

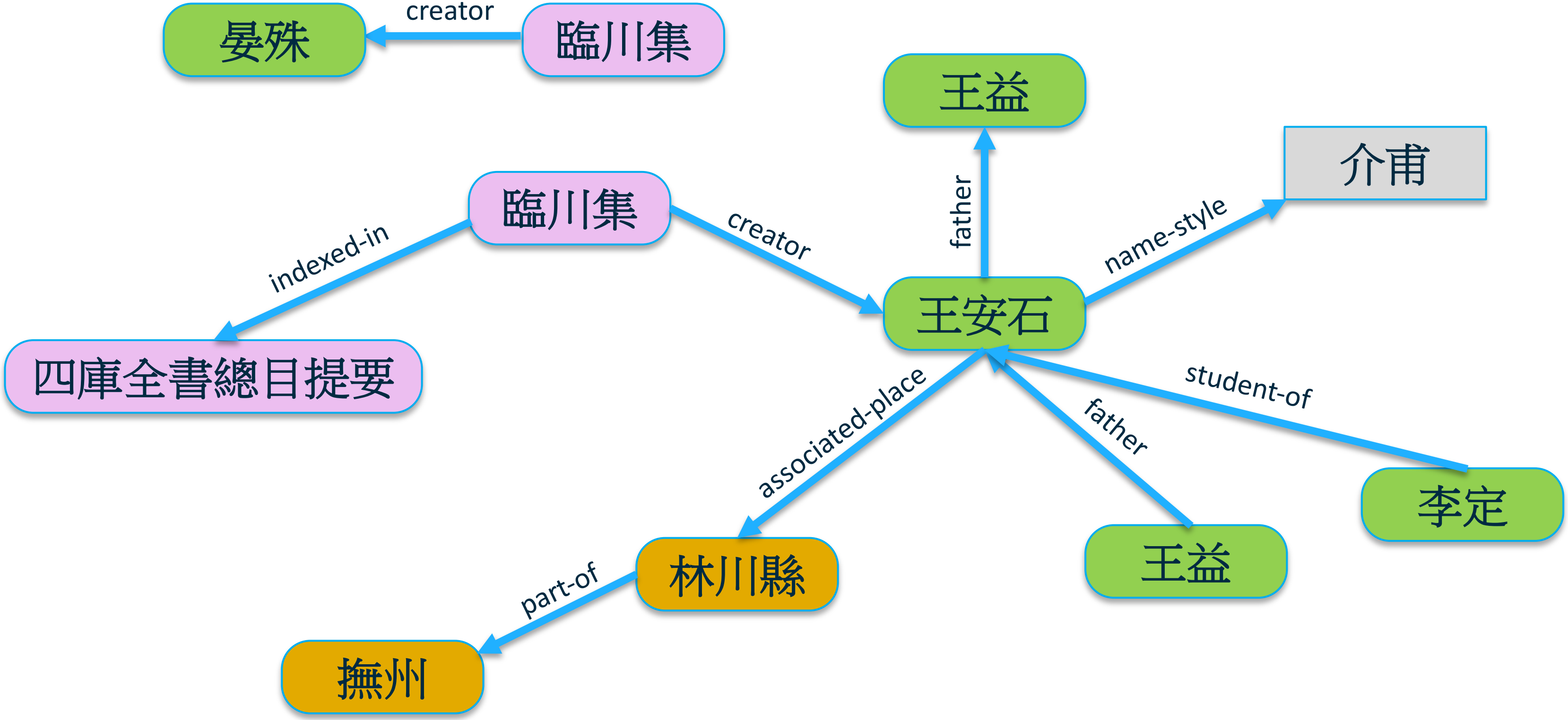
```
PREFIX date: <https://data.ctext.org/date/>
PREFIX cstat: <https://data.ctext.org/statement/>
PREFIX cqual: <https://data.ctext.org/qualifier/>
PREFIX cprop: <https://data.ctext.org/property/>
PREFIX claim: <https://data.ctext.org/claim/>
PREFIX ctext: <https://data.ctext.org/entity/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
```

Defined by W3C

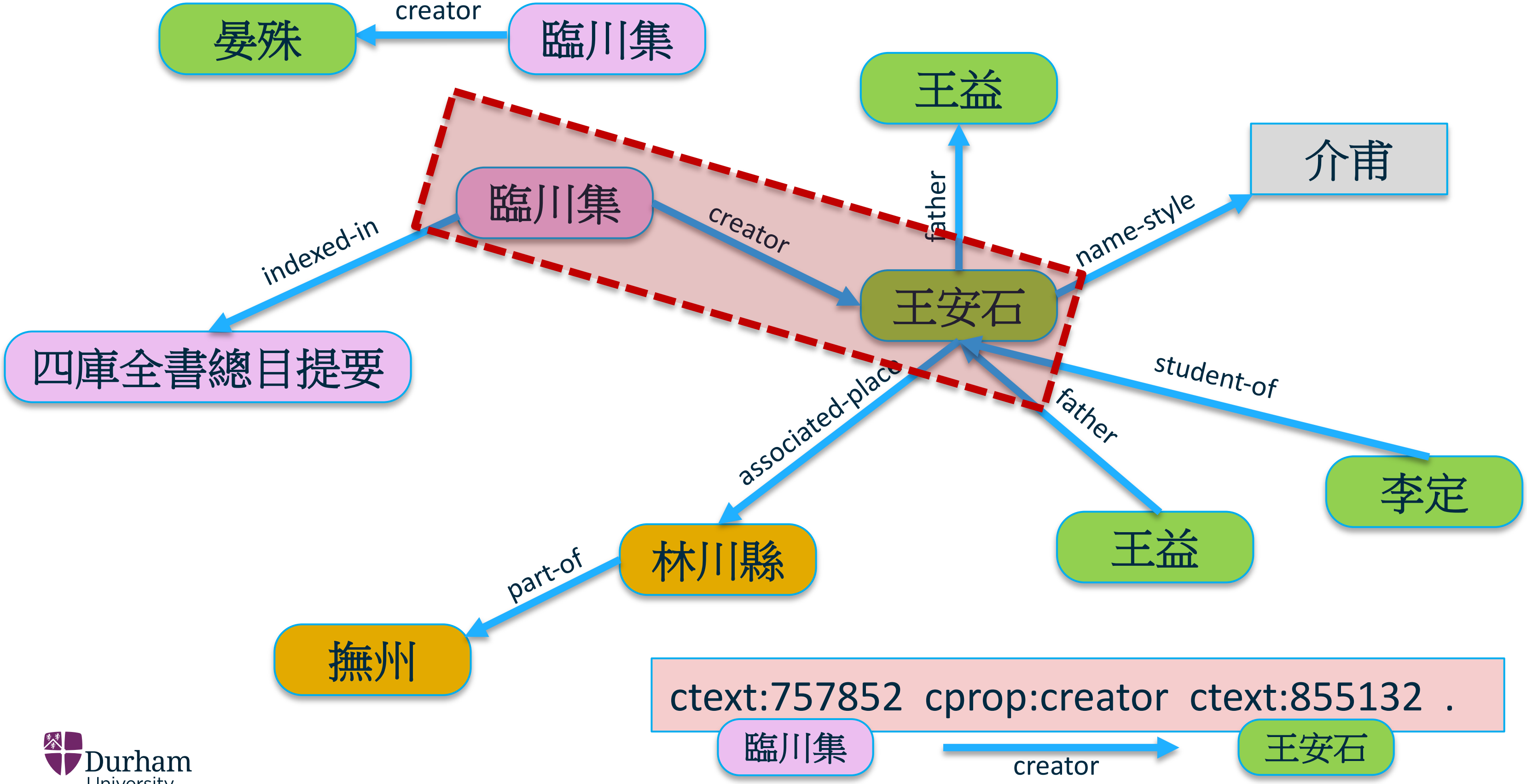
```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX time: <http://www.w3.org/2006/time#>
```

Durham University

RDF data model: representing all data as a graph

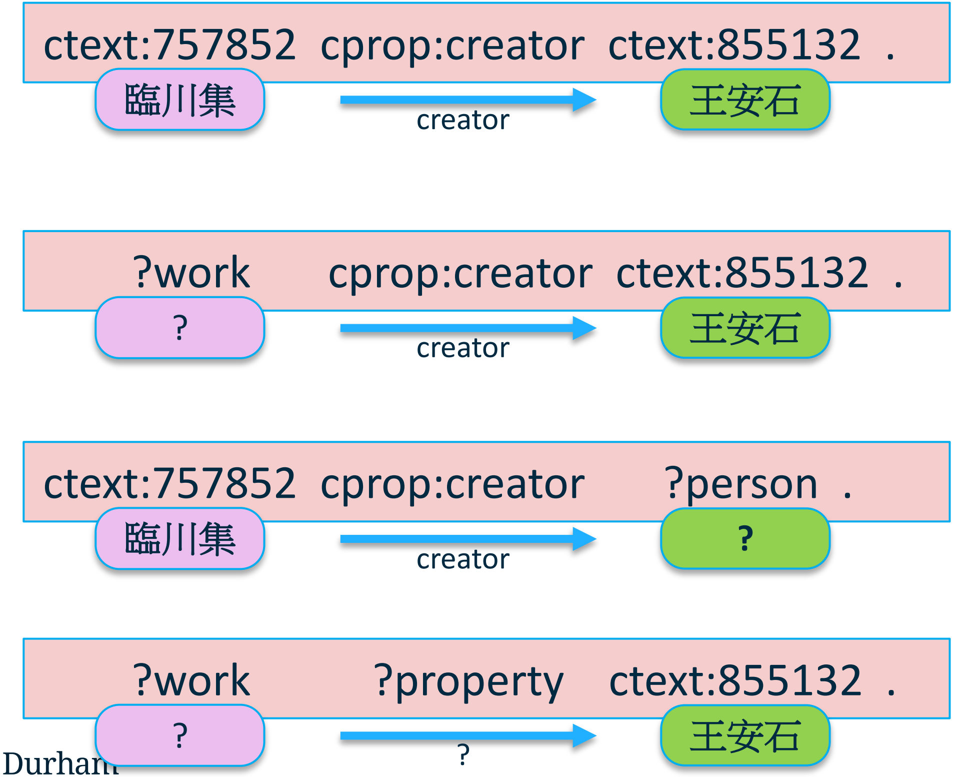


RDF data model: representing all data as a graph



`ctext:757852 cprop:creator ctext:855132 .`

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”

Advanced querying with SPARQL

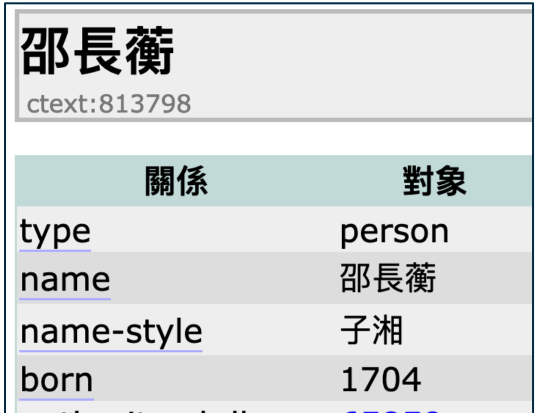
SPARQL is the standard query language for RDF

Somewhat similar to the SQL language for relational databases

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

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

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



邵長衡

ctext:813798

關係	對象
type	person
name	邵長衡
name-style	子湘
born	1704

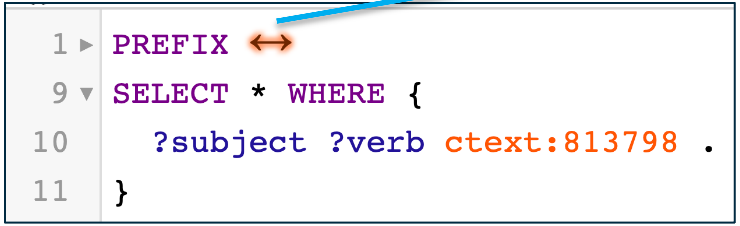
SPARQL, RDF, and prefixes

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

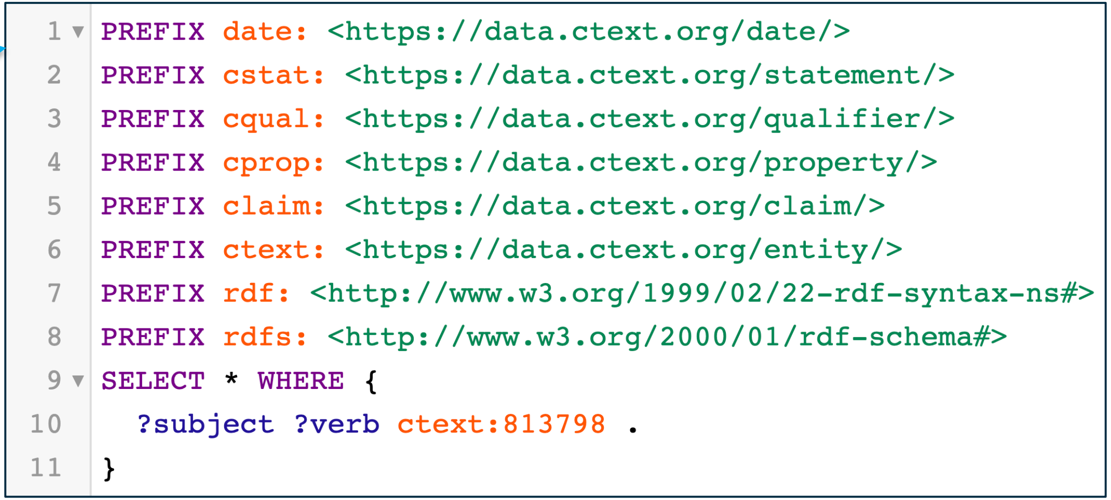
- E.g. https://data.ctext.org/entity/813798

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!



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



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