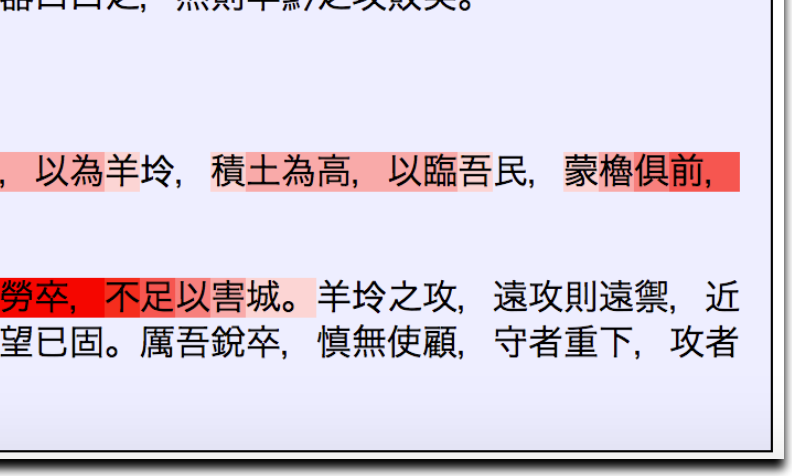
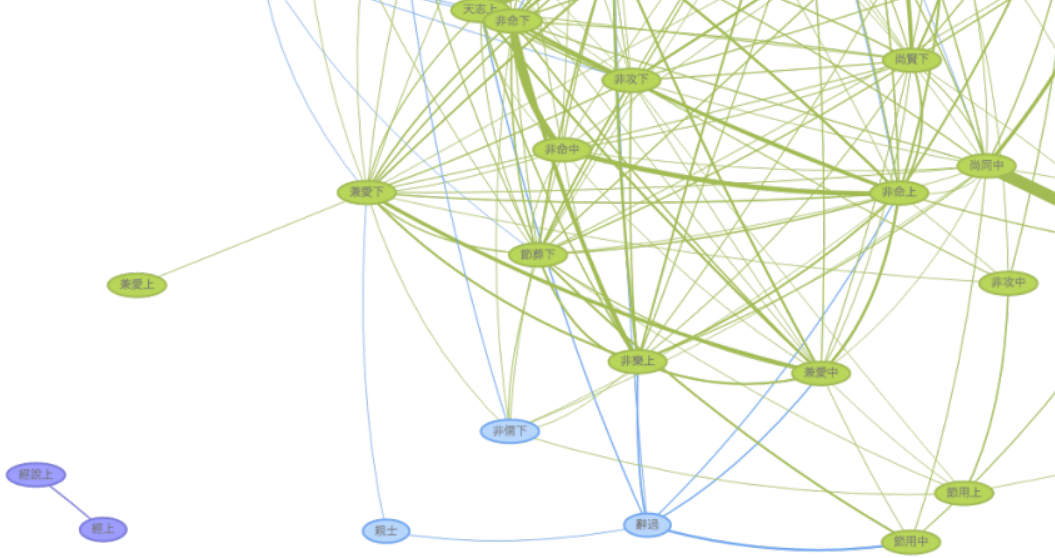
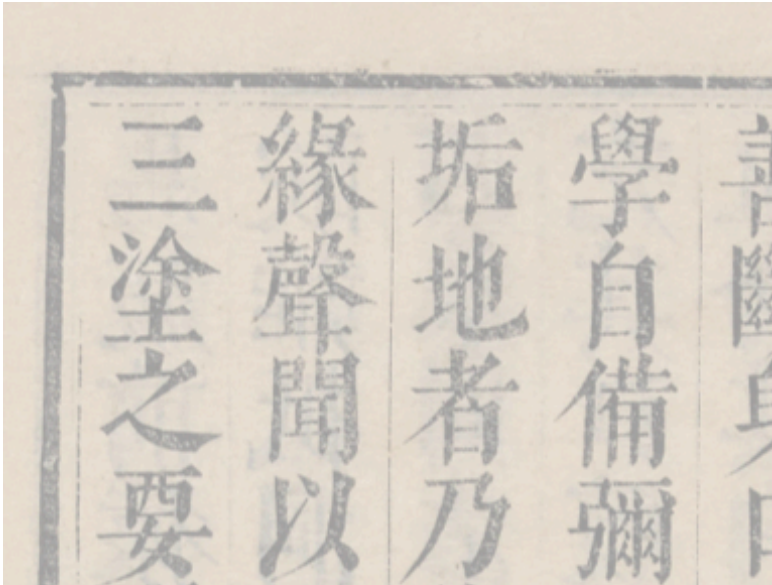




Text mining the Chinese classics

Donald Sturgeon
Department of Computer Science
Durham University
djs@dsturgeon.net

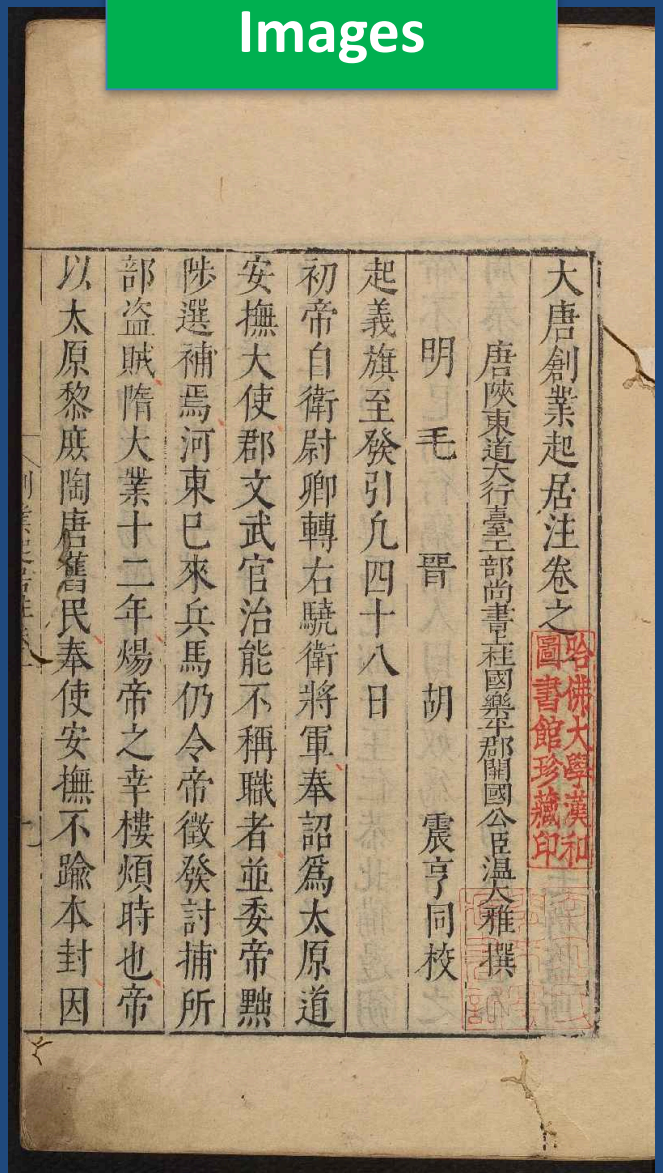


Overview

- General introduction to ctext.org
  - Organization and principles
- Locating and using texts on ctext.org
  - Setup
  - Search
  - Editing
- Digital tools for textual analysis & visualization
  - Text reuse
  - Pattern search (regular expressions)
  - Visualization

Fundamental types of data

Images



Transcriptions

Wiki -> 大唐創業起居注 -> 卷一

《卷一》 [View] [Edit] [History]

1 起義旗至發引凡四十八日

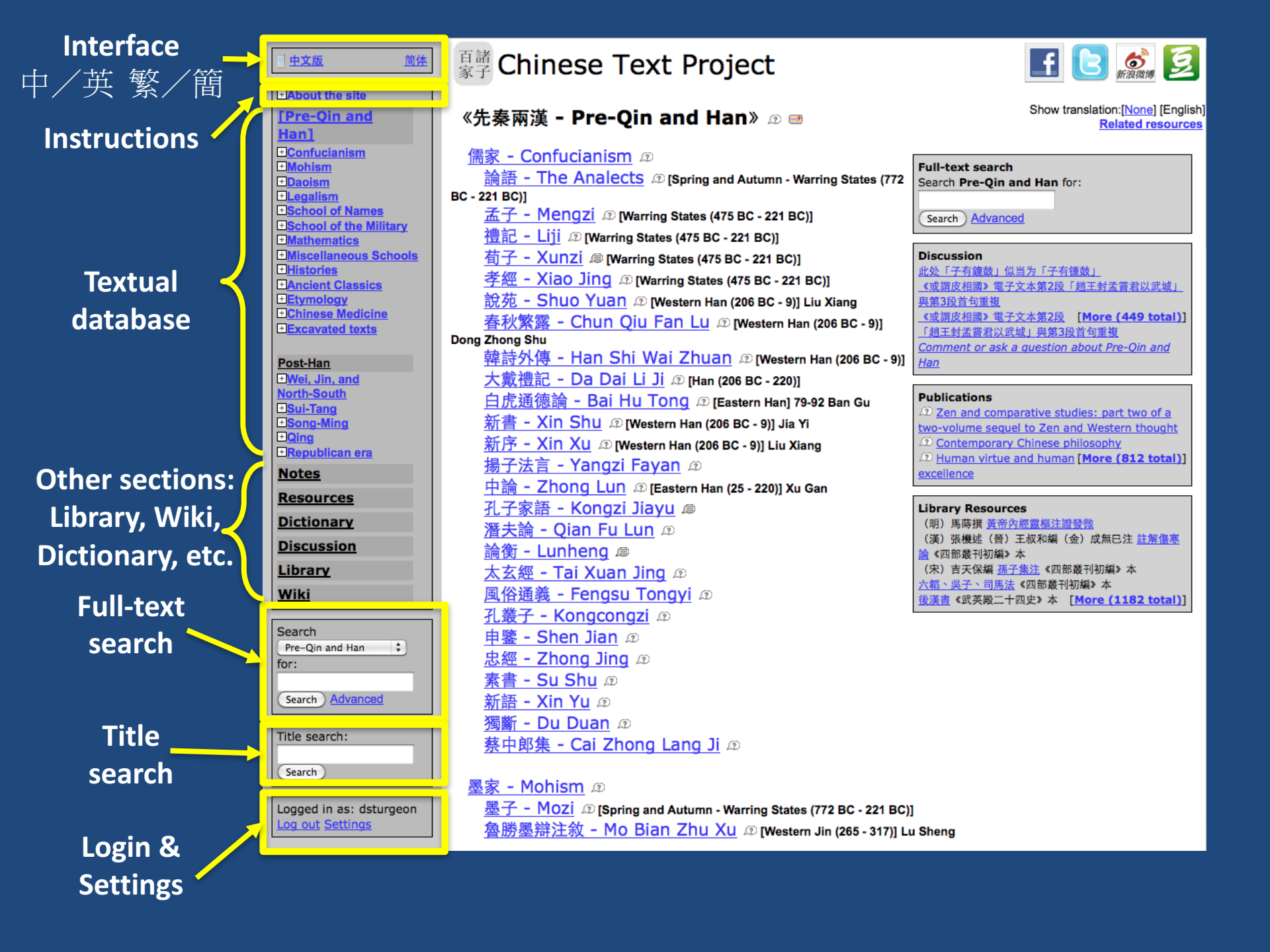
2 初，帝自衛尉卿轉右驍衛將軍，奉詔為太原道安撫大使。郡文武官治能不稱職者，並委帝黜陟選補焉。河東已來兵馬仍令帝徵發，討捕所部盜賊。隋大業十二年，煬帝之幸樓煩時也。帝以太原黎庶，陶唐舊民，奉使安撫，不逾本封，因私喜此行，以為天授。所經之處，示以寬仁賢智，歸心有如影響。

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。帝不得已而行，竊謂人曰：「匈奴為害自，古患之，周秦及漢魏，歷代所不能攘，相為勍敵者也。今上甚憚塞虜，遠適江濱，反者多於蝟毛，群盜所在蜂起。以此擊胡，將何以濟天其或者殆以俾餘。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，不恆其陣。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

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*
    - In particular: primary source materials


Interface
中／英 繁／簡
Instructions
Textual database
Other sections: Library, Wiki, Dictionary, etc.
Full-text search
Title search
Login & Settings
中文版 简体
About the site
[Pre-Qin and Han]
Confucianism
Mohism
Daoism
Legalism
School of Names
School of the Military
Mathematics
Miscellaneous Schools
Histories
Ancient Classics
Etymology
Chinese Medicine
Excavated texts
Post-Han
Wei, Jin, and North-South
Sui-Tang
Song-Ming
Qing
Republican era
Notes
Resources
Dictionary
Discussion
Library
Wiki
Search
Pre-Qin and Han
for:
Search
Advanced
Title search:
Search
Logged in as: dsturgeon
Log out Settings
百諸家子 Chinese Text Project
新浪微博
《先秦兩漢 - Pre-Qin and Han》
Show translation:[None] [English]
Related resources
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 Qiu Fan Lu [Western Han (206 BC - 9)] Dong Zhong Shu
韓詩外傳 - Han Shi Wai Zhuan [Western Han (206 BC - 9)]
大戴禮記 - Da Dai Li Ji [Han (206 BC - 220)]
白虎通德論 - Bai Hu Tong [Eastern Han] 79-92 Ban Gu
新書 - Xin Shu [Western Han (206 BC - 9)] Jia Yi
新序 - Xin Xu [Western Han (206 BC - 9)] Liu Xiang
揚子法言 - Yangzi Fayan
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 - Fengsu Tongyi
孔叢子 - Kongcongzi
申鑒 - Shen Jian
忠經 - Zhong Jing
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:
Search
Advanced
Discussion
此处「子有鐘鼓」似当为「子有鍾鼓」
《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 [More (449 total)]
「趙王封孟嘗君以武城」與第3段首句重複
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)]

Finding Texts

- Left-hand side => “Title search”
- Possible results:

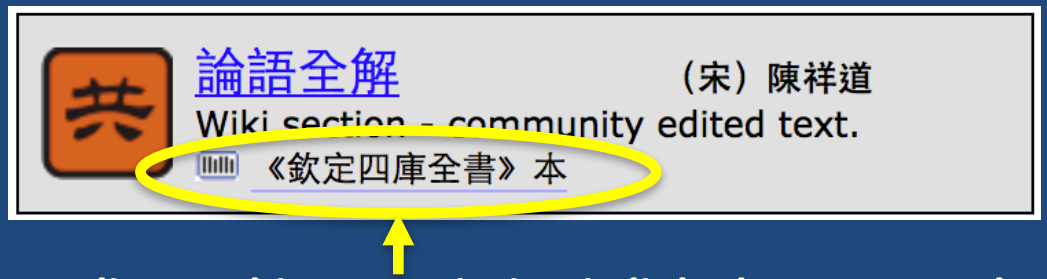
Transcription (text DB)
(not user editable)

Transcription (OCR, wiki)
(uncorrected, editable)

Transcription (wiki)
(user editable)

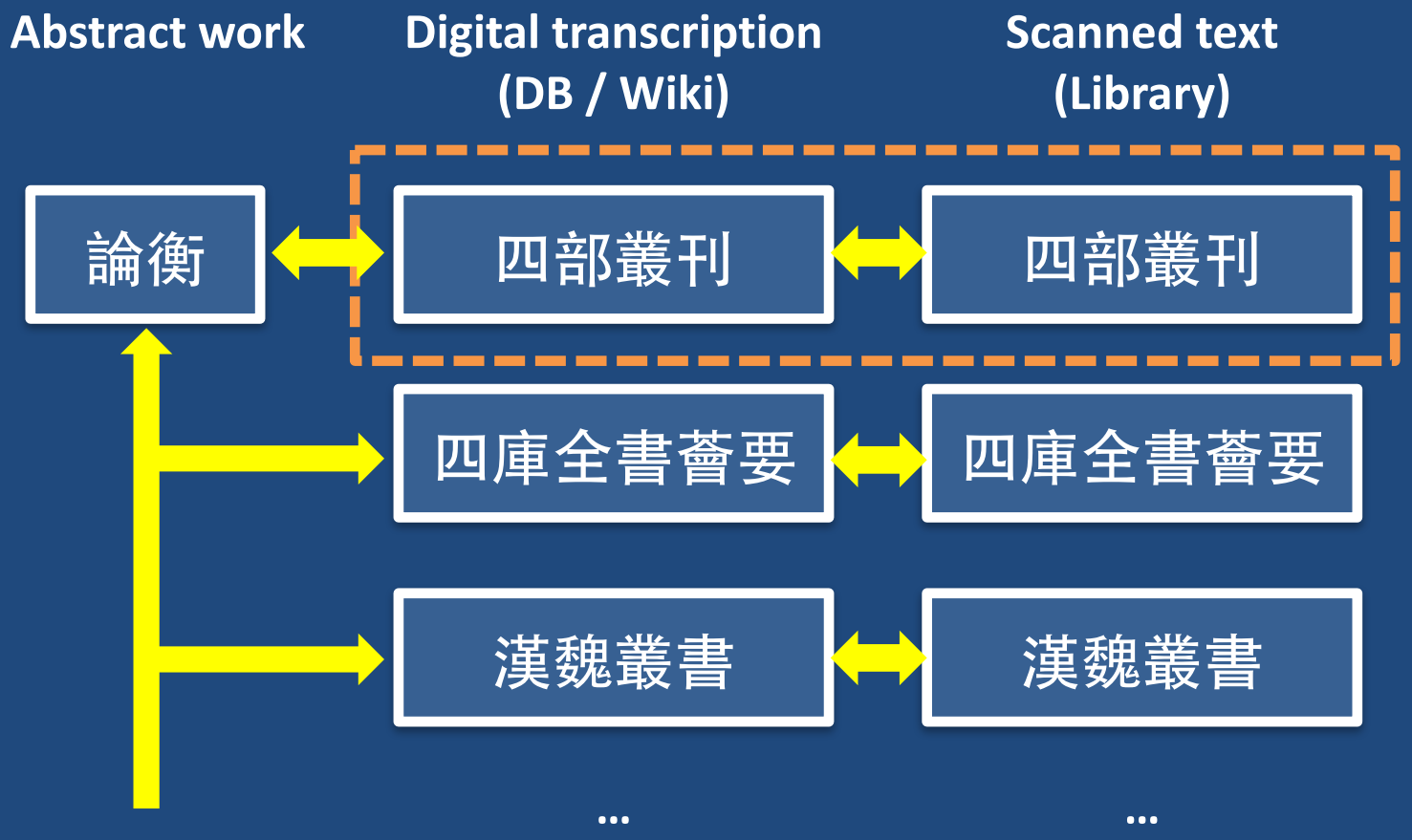
Scanned primary source
(not a transcription)

- Example:



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

Editions


Abstract work
Digital transcription (DB / Wiki)
Scanned text (Library)
論衡
四部叢刊
四部叢刊
四庫全書薈要
四庫全書薈要
漢魏叢書
漢魏叢書
…
…

Hands-on tutorial: Part I

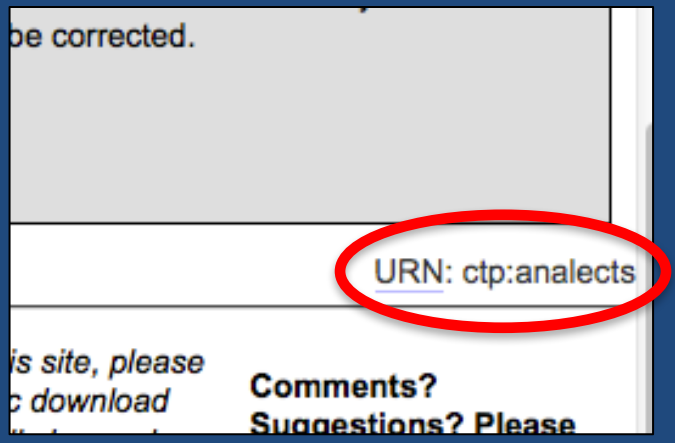
- Overview
  - Setup
  - Finding texts, searching in texts, locating in scans
  - Special functions in the textual database
    - Parallels, translations, commentary
  - Editing
  - Plugins
- (Tutorial: “Practical introduction to ctext.org”)

Hands-on tutorial: Part II

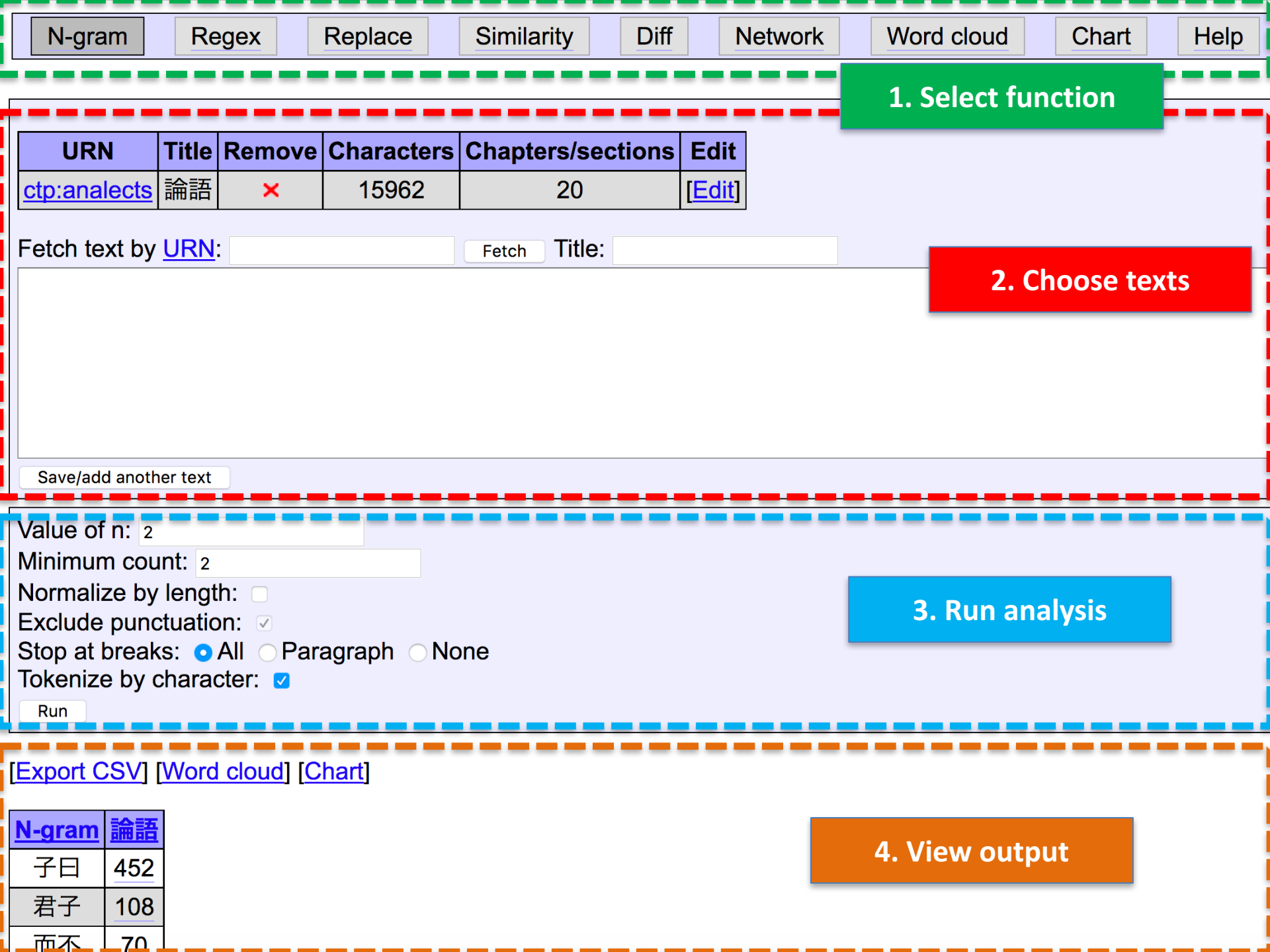
- Text Tools plugin
- Textual analysis tools
  - N-gram counts
  - Text reuse identification using n-grams
  - Regular expressions
  - Cosine similarity
  - Principal Component Analysis
- Visualization tools
  - Network graphs
  - Heat maps
  - Charts
- (Tutorial: “Text Tools for ctext.org”)

CTP URNs

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



- Decoding:
  - 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 output

Hands-on tutorial

- https://dsturgeon.net/maraas