Automation and Collaboration: Exploiting the Digital Medium
自動化與合作：數位媒體的利用

http://ctext.org

Donald Sturgeon（德龍）
Postdoctoral Fellow in Chinese Digital Humanities
Fairbank Center for Chinese Studies
Harvard University
Overview

- Background and introduction
- Scalability, automation, distribution of labor
- Examples
  - Optical Character Recognition (OCR)
  - Application Programming Interfaces (API)
Chinese Text Project (CTP)

- Digital library with digital research tools
- Full-text database of pre-modern Chinese
- Aimed at Chinese and non-Chinese scholars
- Created with limited resources
- Used by tens of thousands of people daily
Texts and data

• Majority of pre-Qin and Han transmitted texts
  – ~6 million characters
• Library of scanned transmitted texts
  – ~13 million pages
• User-editable Wiki
  – ~3 billion characters
• Registered volunteer contributors
  – ~6000 users
Scalability

• To the greatest extent possible:
  – Automate what can be done automatically
    • Automation can often give *approximate* results
      – E.g. OCR-enabled search, machine translation
    • Sometimes *only* automation can provide results
      – E.g. Statistical data on large corpora
  – Crowd-source what cannot be automated
    • Sometimes only human beings will do
      – E.g. Proofreading, interpretation
Optical Character Recognition

• Optical Character Recognition (OCR)
  – Can be applied directly to transmitted texts
  – Not perfect – introduces errors
  – Can still produce useful data
    • Allows fast searching of primary sources
    • Can match an existing transcribed text to a scan
      – Allows fast verification of digital editions
      – Allows us to improve OCR recognition
必近所畏而遠所易。《詩》云：「無棄爾輔，員子爾輻。屢顧爾僕，不輸爾載。」親賢求助之謂也。《貴言》

君子必貴其言。貴其言則尊其身，尊其身則重其道，重其道所以立其教。言貴則身賤，身賤則道輕，道輕則教廢。故君子非其人則弗與之言，若與之言，必以其方。農夫則以稼穡，百工則以技巧，商賈則以貴賤，府史則...
必近所畏而遠所易。《詩》云：「無棄爾輔，爾予爾輻。」親賢求助之謂也。

《貴言》

君子必貴其言。貴其言則尊其身，尊其身則重其道，重其道所以立其教。言費則身賤，身賤則道輕，道輕則教廢。故君子非其人則弗與之言，若與之言，必以其方。農夫則以稼穡，百工則以技巧，商賈則以貴賤，府史則...

Scanned image + Merged: accurate, linked, punctuated
OCR – Training with real data

• Identify concrete instances of *each* character
• Train OCR engine to recognize these forms natively

2015 DADH
Crowd-sourcing

• OCR results are imperfect
  – Still useful – enable full-text search
• Users can be active contributors
  – Frequent users *want* to correct mistakes
  – Corrections are easily validated
  – Records are kept of who edited what
  – Previous versions can be easily restored
Making corrections
Making corrections

- Changes propagate to all visualizations
  - Search index automatically updated
Application Programming Interfaces

• Formalize automated exchange of data
  – Standardize procedures for communication
  – Perform complex functions on externally supplied data

• Allow distribution of labor
  – E.g. Google Maps API

• Allow tools to work intuitively together
Third-party APIs on CTP

- OpenSearch API
  - Defines machine-readable search interface
  - Adds ctext.org search function to web browsers
  - Allows federated search (e.g. CrossAsia)
Third-party APIs on the CTP

• Pingback (trackback) API
  – Notifies a website when another site links to it
  – Links to CTP uniquely identify passages of text
    • http://ctext.org/analects/wei-zheng#n1119
      => 子曰：「詩三百，一言以蔽之，曰『思無邪』。」
    • http://ctext.org/dao-de-jing#n11631
      => 反者道之動；弱者道之用。
  – Discussion can be linked back from cited passages
Pingbacks as Textual References

I don’t find the last explanation convincing, I believe that the *xin* 辛 + *kou* 口 explanation is the right one since we clearly find these components in the early forms of *yan* 言. Gu Yankui 谷衍奎 his explanation might be inspired by the *Erya* 羹雅 *dictionary* which says that a *yan* 言 is a name for a large flute.

On oracle bones *yan* is ... *sound* is also used as the early form of *xin* 欲, meaning ‘to offer/entertain’. This meaning was also used in later bronze inscriptions (陳初生, ‘金文常用字典’, p. 239).

http://ctext.org/er-ya/shi-yue#n38388

2015 DADH
Chinese Text Project API – Goals

• Allow other projects access to CTP data
  – Avoid having to maintain duplicates of *all* data
  – Obtain the current version in real time
• Allow two-way linking with the CTP
  – Allow corrections and improvements to be pooled
  – Allow easy access to scanned editions, parallel passages, linked commentaries etc.
CTP API – Two-way communication

• Access other resources from within the CTP
  – Plugin model
    • User-definable specification of how to pass data
      – Send a chapter of text to an external tool
      – Look up a character or word in another dictionary
    • Simple definition in XML

• Access CTP data/tools from within other tools
  – Structured data exchange
    • Request CTP textual data, metadata, statistics, ...
筆

6—6 strokes = 12 strokes total.

related glyphs

502 a—b. *bǐwèi* / bǐèi / yù Shuowen says: writing stylus or pencil (no text); loan for id. following, thereupon (Shi); a particle (Shi). Gl. 762. b. is pre-Han (inscr. 451, name). The graph is a drawing.

503 a. *bǐ* / bǐ / *pí* writing pencil (Li).

504 a—b. *bǐ* / bǐ / *pi* writing pencil (Li).

2015 DADH
CTP API – Textual plugins

[Image of a screen showing various tools and options related to text analysis and frequency comparison.]

Pre-Qin and Han → Daoism → Dao De Jing

《道德經 - Dao De Jing》

Full text export

URN for text

2015 DADH
Further Information

- **English:**
  - http://ctext.org/introduction
  - http://ctext.org/instructions/ocr
  - http://ctext.org/faq

- **Chinese:**
  - http://ctext.org/introduction/zh
  - http://ctext.org/instructions/ocr/zh
  - http://ctext.org/faq/zh

- **API etc.:** djs@dsturgeon.net