Harnessing Digital Sources for Sinology Research

Jieh Hsiang (項潔)
National Taiwan University

Naïve diagram of humanities research
What has changed in the digital world?

 Resource over the web

 Motivate → Gather → Process → Organize → Analyze → Visualize → Interpret

 Publish

How do most humanities scholars cope with the change?

• Draw research material from the Web (through search and retrieve)
• Then work on them pretty much the conventional way
What is the problem?

• There is an overwhelming amount of information available, it is often impossible to work with them manually

• Without digital tools, we can’t effective harness digital contents

But we are using tools already...

• Obtaining content
  • Web browser
  • Search engine

• Organizing content
  • Excel
  • Endnote, Zotero, Mandeley

• Social network Visualization
  • Delphi
  • Palladio
  • ArcGIS
  • etc
What is missing with using the tools?

- There should be **tools for every step** of the research process (except interpretation, which is inherently a human effort) in an integrated fashion
- Some tools are hard to learn (especially GIS tools)
- Tools for **analysis are underutilized**
- No tools to help with access once the materials are downloaded
- No fluid **interaction** between content and tools
- No effective method to **connect** humanities scholars and tool developers

---

Digital resources = digital contents + digital tools

Two main subjects in this talk
- Connecting digital contents and digital tools
- Connecting humanities scholars and IT researchers

Assumptions
- Contents are already in digital form
  - Going from analog to digital is important, but that’s a different issue
- Work mainly with Chinese texts, not images
  - The principle is the same, but tools are different
The digital library approach to develop scholarly content

- Bypass individual research problem/motivation
- Data Collection
  - Comprehensively and systematically collect material of a specific domain (usually with institutional support)
- Data Processing
  - Organize and digitize material according to an SOP
- Data Access
  - Build a large scale digital library system
  - Access through browse and keyword search

The digital library framework

- Focus on context collection, process, and access
- Instead of simple organization, build a database to access digitized material!
From DL to DH

- DL focuses on building large scale, high quality digital repositories
- DH also emphasizes on incorporating analytical and visualization tools into the DL
- DH system: Integrate content and tools within a system to perform context discovery (exploring relationships among digital contents)

Example: ctext.org

- Full-text Chinese classics developed by Donald Sturgeon
- Example shows the ability to find similar text in other books
Example: CBETA

- Full-text system of Buddhist canons built at Chinese Buddhist Electronic Text Association
- Example shows dates of books in which keyword appears
- All canons must have the era when it was translated

Ex: THDL (Taiwan History Digital Library)

- Built by Research Center for Digital Humanities of NTU
- (Tagged) full-text of primary historical material of Taiwan, focusing on Qing dynasty
- Emphasize context discovery and analysis
- In addition to full-text search, provide explainable textual contexts (relations among documents) of a search result set

- RCDH has built over 30 large scale context-discovery systems on different historical materials and Chinese classics
  - Accessed by more than 25,000,000 sessions. Cited in >1,000 papers
Metadata contexts

- Contexts implicit in the metadata
- Chronological distribution, geographic distribution, authors, sources,

Statistical contexts

- Contexts from statistical analysis of texts
- Co-occurrence of terms, term1 vs term 2, prefix/suffix analysis, etc
Semantic contexts

- Contexts that are domain-specific, usually require pre-computation
- Land transitivity graph (for land deeds), IE/M (imperial edicts/memorials) diagram

Context-discovery systems – pros

- Through extensively collecting of material on a specific domain, allow scholars to have systematic access of related research material
- Through context analysis, provide a bird-eye’s view of a document set and basic contexts among documents
- Give a user an effective way to utilize retrieved documents
- Often lead to new discoveries of problems
Context-discover systems - shortcomings

• Contents and tools of a large DH system are tightly coupled
  • The database is read-only (can’t easily add new material)
  • New functions can only be added by experts (mainly the designer of the system)
• But scholars usually have their own research focus and personal collection of related material
• Personal material cannot be incorporated into an existing context-discovery system nor can utilize the tools in the system
• In general, the personal need of the research cycle cannot be satisfied by using a static context-discovery system

How to incorporate a user’s own material into a large DH system?

• You can’t
How to satisfy a humanities scholar’s individual IT need?

A DH platform for individual scholars to work in the digital environment

- Data collection
  - Tools to access/download relevant research material available on the Web
- Data processing
  - Tools to process (own and downloaded) material into proper form
- Data access
  - Simple way to build a personal DL to provide easy search and access of personal material
- Data analysis
  - Tools to analyze material and to find textual context
- Data visualization
  - Tools to provide visual presentations of analytic results
DH platform for individual scholars

System versus Platform

- Contents are fixed. Can’t add personal content
- Contents and tools are tightly coupled
- Contents and tools can only be added by the administrator
- Contents can be added/removed/modified, accessed, used
- Tools can be added independently
- Contents and tools can interact freely
- All users can add own content or tools and make them available to others
DocuSky: A personal DH platform

• Developed at RCDH, NTU
• Chief architect: Dr. Hsieh-Chang Tu
• Project manager: Dr. I-Mei Hung, Dr. Chijui Hu (co-MP)
• An attempt to realize the scenario given above
• Connecting (personal and Web) contents and tools
• Connecting scholars and IT developers

Division of work

Scholar
• Initiate research problem
• Gather related material
• Interpret results

DocuSky
• Obtain additional material over the Web
• Tools for processing material
• Create a searchable database of processed content
• Tools for analysis
• Tools for visualization
The design principle of DocuSky

- DocuXML: a standard to connect contents and tools
- One-click personal database
  - With just one click, turn any document set in DocuXML format into a full-text searchable database with metadata context post-classification
- Utilize open access content
  - We don’t create own content, unless necessary
- Utilize open access tools
  - We don’t create own tools, unless necessary

How to connect external tools?

- To ensure the compatibility between existing tool’s own format and DocuXML, sometimes need to write a converter if the tool uses a different format
- So far support csv, .xlsx (MS Excel), .txt, .html (Markus format), DocuXML
An overall picture

- Online text repositories
  - CBETA
  - CTEXT
  - Kanripo

- Personal data
  - Text files
  - MARKUS html (tagged text)
  - Excel metadata

How to access external content

- DocuXML API in
- API out

- Search and Browse
- DocuGIS
- Tag/Term Stats Tool
  (Graphed by Palladio)

Personal DH Platform
- Motivate
- Gather
- Process
- Access
- Analyze
- Visualize
- Interpret
- Publish
How to connect open access content?

- Each open repository needs to be accessed separately, mainly through API
  - Ctext.org
    - Obtain plain text through API
  - Kanripo
    - Obtain plain text with simple metadata through API
  - CBETA
    - Obtain files already in DocuXML form through API (thanks to Cbeta team)
  - THDL
    - THDL sends files to DocuSky in DocuXML format
  - Wikisource
    - Coming soon...

Tools for processing/annotating content
Annotating tool(s): MARKUS

• Markus, developed at Leiden by Hilde De Weerdt and Brent Ho
• Automatic markup of person names (using CBDB and others), place
  names (using CHGIS and TWGIS), temporal references (Dila)
• Keyword markup
  • User needs to create metadata
    attribute and list of terms
• Manual markup
• Comparativus
  • Comparing two texts
  • (Recently added feature)

The importance of tagging

• Chinese lacks delimiter (such as space) to naturally separate words
• Tagging not only identifies meaningful words, but also enriches the
  usefulness of the text
• By using authority files such as CBDB, Markus enables the potential of
  linking texts to other tools (such as GIS) and semantic queries
As a remark, when digitizing, we should...

- Build metadata
- Produce full-text when possible
- Include image when possible
- When possible, tag person/place names, normalize dates, tag objects, identify geographic coordinates

An example of Bencaojin.jizhu (本草經集注)

- Obtain full-text of BCJ.JZ from Kanripo
- Use Markus to tag person name (with CBDB) and place names with coordinates (with TGAZ)
- Design additional tags such as #DrugName, #DrugProperty, #DiseaseName, #Harvesting, which with a list to tag by Markus automatically
For accessing the prepared content: build a searchable database with one-click

Building a searchable database of DCJ.JZ

- Convert Markus result (.html) to DocuXML (a couple of clicks)
- One click to build database
The BCJ.JZ DH (context-discovery) system

- [Link](http://docusky.org.tw/DocuSky/webApi/webpage-open-3in1.php?db=%E6%9C%AC%E8%BD%89%E7%86%99%E9%86%83%E8%A6%9F&corpus=%E6%9C%AC%E8%BD%89%E7%86%99%E9%86%83%E8%A6%9F&queryBase=.all&spType=tagAnalysis&snippets=null)

### BCJ.JZ system post-classification using #Drug_Actions

- [Screenshot](image-url)
What drugs prolong life (#MainDrugName)?

Tagging cloud of drugs that prolong life
Geographic sources of drugs that prolog life

Using DocuGIS plugin to show geographic distribution

Connecting CBDB and DocuGIS
What books were used in bcj.jz?
Using suf-pref query: “< xxxxxxxxxx >”

Sampling of tools (term-stat and DocuGIS)
Analytic tools in DocuSky

- Discovery tools
  - Word-clipping
  - Document recommendation
  - Text comparison (forthcoming)

- Statistical tools
  - Word frequency
  - Tag/term stats tool: comparing word occurrences among documents

Tag/Term Stats Tool (Bo-yu Hsieh)

- Purpose: word frequencies in documents
- Given texts and a list of terms, generate frequencies in several formats to be used by other tools
Simple example of Term StatsTool

- Texts: Taiwanese news (9,497 reports)
- Terms: common idioms (5,152 idioms)

<table>
<thead>
<tr>
<th>Idioms</th>
<th>f/news</th>
<th>times</th>
</tr>
</thead>
<tbody>
<tr>
<td>不可思議</td>
<td>68</td>
<td>74</td>
</tr>
<tr>
<td>流離失所</td>
<td>43</td>
<td>45</td>
</tr>
<tr>
<td>屢出不窮</td>
<td>43</td>
<td>44</td>
</tr>
<tr>
<td>史無前例</td>
<td>40</td>
<td>44</td>
</tr>
<tr>
<td>雪上加霜</td>
<td>38</td>
<td>40</td>
</tr>
<tr>
<td>捲土重來</td>
<td>37</td>
<td>40</td>
</tr>
<tr>
<td>墓葬之災</td>
<td>37</td>
<td>39</td>
</tr>
<tr>
<td>軒然大波</td>
<td>37</td>
<td>37</td>
</tr>
<tr>
<td>手無寸鐵</td>
<td>33</td>
<td>35</td>
</tr>
<tr>
<td>個別可危</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>虛無戒計</td>
<td>31</td>
<td>32</td>
</tr>
<tr>
<td>一朝而行</td>
<td>26</td>
<td>28</td>
</tr>
</tbody>
</table>

Frequencies of words in corpuses

- Corpuses: Quantangshi (全唐詩), Quansongci (全宋詞)
- Lists of terms: Beasts and Birds from Taipinyulan (太平御覽)

<table>
<thead>
<tr>
<th>Term</th>
<th>Full Tang</th>
<th>Full Song</th>
</tr>
</thead>
<tbody>
<tr>
<td>鳥</td>
<td>4396</td>
<td>47957</td>
</tr>
<tr>
<td>鳥</td>
<td>1234</td>
<td>18932</td>
</tr>
<tr>
<td>鳥</td>
<td>931</td>
<td>3850</td>
</tr>
<tr>
<td>鳥</td>
<td>710</td>
<td>1226</td>
</tr>
<tr>
<td>犭犬</td>
<td>540</td>
<td>197</td>
</tr>
<tr>
<td>犭犬</td>
<td>405</td>
<td>812</td>
</tr>
<tr>
<td>羽羊</td>
<td>400</td>
<td>164</td>
</tr>
<tr>
<td>羽羊</td>
<td>332</td>
<td>123</td>
</tr>
<tr>
<td>鳥</td>
<td>230</td>
<td>76</td>
</tr>
<tr>
<td>鳥</td>
<td>1266</td>
<td>1099</td>
</tr>
<tr>
<td>鳥</td>
<td>995</td>
<td>910</td>
</tr>
<tr>
<td>鳥</td>
<td>33</td>
<td>535</td>
</tr>
<tr>
<td>虎</td>
<td>230</td>
<td>458</td>
</tr>
<tr>
<td>虎</td>
<td>230</td>
<td>428</td>
</tr>
</tbody>
</table>

- Importance of Apes (猿) dropped significantly (from 2.57% to 0.85%)
- Importance of Swallow (燕) increased significantly (from 2.93% to 9.53%)
Visualization of co-occurrence of terms in corpuses (Hsieh-Chang Tu)

• Given several document sets (corpuses), and a list of terms
  • Question: what are the common terms among two or more corpuses?
  • What terms only appear in one corpus?

Steps
• Prepare corpuses/term list
• Feed them to term stats tool
• Export the output in csv to Palladio (Stanford)
Example: Daoist deities in two novels (Dr. Chijui Hu)

- **Question**: comparing Daoist deities in *The Journey to the West* (西遊記) and *The Apotheosis Tales* (封神演義)
- **Why**: the former was written by Wu Cheng-En (吳承恩), the latter is unknown. People suspect that Wu also wrote the latter. Perhaps we can see something by comparing the deities that appear in both?

**DocuSky scenario**

**Preparing the material**
- **Step 1**: Download the full texts of the novels from ctext.org through API
- **Step 2**: Obtain a list of deities from the Web
- **Step 3**: Prepare the novels in DocuXML format (easy with the converter in DocuSky)

**Generate the graph**
- **Step 4**: Input the two texts and the list into TCG (the Term Co-occurrence Graph tool)
- **Step 5**: Export the result in csv format
- **Step 6**: Feed the csv into Palladio
  - **Viola!** Step 3 to Step 6 less than an hour

---

**Daoist deities in JtW and TAL**

- Many minor deities in common
Visualization tools in DocuSky

• Geographic tools
  • Geoport
  • DocuGIS

• Social network tools
  • Developed at Stanford

• Temporal tools
  • DocuTime (forthcoming)

• Comparative reading of several texts

Why DocuGIS?

• I have a text and I want to know the locations of the places mentioned, but GIS can only be done by experts, right?
• Even if I can plot the points, can I interact with it?
• Can I see the locations in its original historical setting?

DocuGIS (Dr. Nung-yao Lin)

• Easy to learn. Everyone can do it
• Provide interaction through full-text search
Travel route of Xuanzang according to *Pilgrimage to the West during Tang Dynasty* (大唐西域記)

- Full-text of PtWdT
  - Download from ctext using DocuSky API
  - (PtWdT describes the same travel as the fiction *The Journal to the West* 西遊記, but was written by the monk Xuanzang himself)
- Geographic locations
  - Dharma Dram location authority (http://authority.dila.edu.tw/place/)
- Build Excel file of names and locations of places and the full-text
- Obtain csv from the Excel file
- Upload to DocuGIS and done!

Making the Excel of PtWdT

One click to turn Excel into csv
Paste csv to DocuSky, then choose upload

Searchable GIS system of PtWdTD completed!
How to plot the route?

- Use Index as order, and choose “route” in the reference map option

One click and the route is complete
Interaction with user: query with Brahman (婆羅門)

Replacing the map by a Tang map (map courtesy of Academia Sinica)
Connecting scholars and tool developers

- Through DocuXML and other common formats such as csv

Many tools are motivated from requests of scholars

Who built DocuSky?

Built by full-time staff
- The platform architecture
- DB-builder
- DocuXML standard
- DocuWidgets
- Term-stat tools
- Some converters

Built by students and external
- Most converters
- Dialogue annotator
- Batch tagging tool
- Metadata attachment tool
- DocuGIS (and other GIS tools)
- DocuTime lite
- CompareReader
- Markus (Leiden)
- APIs (Cbeta API done by DILA)
- Palladio (Stanford)
The DocuSky experiment so far

- Over a thousand registered users
- Corpuses used by scholars range from Ying Oracles to elementary textbooks, at least 100 cases
- Over 15 workshops/tutorials targeting conducted at China, Singapore, U.S. and Taiwan

DocuSky: a digital platform on which a scholar can do all the work by herself
Concluding remarks

• We advocate the division of work between a scholar and digital tools, and that effective utilization of tools is the only way to harness the complexity raised by digital contents.

• To do so we propose a new concept of platform with which new contents and new tools can be incorporated easily.

• DocuSky aims at becoming such a platform where a scholar can work with digital contents and tools without loosing independence.

• DocuXML provides a way for tools and contents (and their developers) to participate and interact.

• A lot more need to be done...

Thank you
http://docusky.org.tw
http://www.digital.ntu.edu.tw