PANEL 300. C  Saturday 5:00pm-7:00pm Gregory A/B, 2nd Level

WORKSHOP:
Computational Tools and Digital Resources for Chinese History and Literature

Michael A. Fuller
University of California, Irvine

From Texts to Databases:
The Computer Modeling and Analysis of Historical and Literary Materials
First Question: For what sort of historical or literary material do you want to develop a model?
Next Question: Why? What do you want to do with the data?

For example, let us consider letters written from one person to another. There are many reasons to study letters, some more amenable to computer analysis than others.

- We could be interested in **content**:  
  - What are the topics?  
  - What people are mentioned  
  - What places or events or texts are mentioned

- We could be interested in **stylistic and rhetorical issues**:  
  - What levels of language are used?  
  - Are there particular repeated persuasive strategies?  
  - How do the authors present themselves and their recipients?
- We also could be interested in the sociology of production and circulation of letters:
  - Are there discrete networks of people writing one another?
  - Do these networks have particular characteristics in terms of geographic distribution or social status?

How we model a “letter” depends on what we want to look at. **BUT:** a prudent design preserves “hooks” so that we do not need to repeat work later.
This requires thought and planning.
Basic Entity: **LETTER:**

- *Author (Person)*
- *Recipients (List of People)*
- *Date of composition*
- *Place of composition*
- Text of Letter

For sociological analysis of a group of letters, we might want to know various facts about the people involved in the letter: place of origin, status, age, birth date, and so on.

Because we do not want to repeat that information for authors and recipients in the record for each letter, we create a separate entity:

**PEOPLE:**

- *Name*
- *Place*
- *Birth date*
- *Status, etc.*
We also notice that there are other categories of information for **LETTERS** and **PEOPLE** that are complex and in fact should be considered **entities**:

**PLACES:**
- Name
- Type (prefecture, county, etc.)
- x-coordinate
- Y-coordinate

**STATUS:**
- Label
- Metrics for status (date of jinshi, official title, etc.)

Now we can think about how these various entities interact:
Flexible, Extensible Design allow one to approach interactions from many angles
Does Status shape Rhetoric?
Does Place shape Rhetoric?
A Preliminary Model:
The 8,000 Letters (so far) in the *Complete Song Dynasty Prose*

It started as an Excel Spreadsheet of the Titles, Authors, Recipients:
The spreadsheet, imported into Access, becomes a table:

Note the structure of the data:
The letters have IDs
The Authors and Recipients have IDs: these are the CBDB IDs
Because the people use CBDB IDs, I can import CBDB tables for such entities as: PEOPLE, PLACE, KINSHIP, and ASSOCIATION (and can import more if needed later)
However, starting from a spreadsheet can lead to bad design because one tries to model the data as single records in a flat table.

The spreadsheet I was given, for example, had only one column for the recipient. What happened when a letter had more than one addressee? They created a new record (and a new letter ID) for each recipient:

<table>
<thead>
<tr>
<th>LetterID</th>
<th>AuthorName</th>
<th>AuthID</th>
<th>LetterTitle</th>
<th>LetterSup Text</th>
<th>Recipient Name</th>
<th>RecipientID</th>
<th>VolNum</th>
<th>PageNum</th>
<th>LetterNo tes</th>
<th>LetterType</th>
</tr>
</thead>
<tbody>
<tr>
<td>51652</td>
<td>黃庭堅</td>
<td>7111</td>
<td>與洪氏四甥書</td>
<td>一</td>
<td>洪羽</td>
<td>10618</td>
<td>105</td>
<td>108</td>
<td></td>
<td>書</td>
</tr>
<tr>
<td>51653</td>
<td>黃庭堅</td>
<td>7111</td>
<td>與洪氏四甥書</td>
<td>二</td>
<td>洪炎</td>
<td>10616</td>
<td>105</td>
<td>108</td>
<td></td>
<td>書</td>
</tr>
<tr>
<td>51654</td>
<td>黃庭堅</td>
<td>7111</td>
<td>與洪氏四甥書</td>
<td>三</td>
<td>洪朋</td>
<td>10615</td>
<td>105</td>
<td>109</td>
<td></td>
<td>書</td>
</tr>
<tr>
<td>51655</td>
<td>黃庭堅</td>
<td>7111</td>
<td>與洪氏四甥書</td>
<td>四</td>
<td>洪芛</td>
<td>10617</td>
<td>105</td>
<td>109</td>
<td></td>
<td>書</td>
</tr>
</tbody>
</table>
“Recipients” is a classic **one-to-many** relationship: **one** letter may have **many** addressees. To allow us to record these addressees, it is best to create a separate table:

**RECIPIENTS**

| Letter ID | Recipient ID | Recipient Place | Receipt Date | etc. |

Note that there is no point repeating any information about the letter or about the recipient in this table: we have that information elsewhere in the database.

The principle of not repeating information—that is, recording it just once in the database—is called **normalization**.

(You’ll note, however, that the LETTERS table violates this principle in recording people’s names.)
What can we do with the data at present?
First, we can sort out the data a bit:

Among the 8,004 letters, there are 2,793 sets of authors and recipients. There are 1,947 individuals listed as authors or recipients.
We also can look at NETWORKS of letter-writers: was everyone writing to everyone else or were some individuals at the center of the networks?
If we look at a fifty-year period, 1076 to 1125, determined by author index year:
Looking in details, we get:
If we repeat this for 100 years later (1176-1225), we get the components:
We find the “usual suspects” at the center:
If we look at how many people were writing to relatives, we get a slightly surprising result: