*Jeffrey Tharsen tharsen@uchicago.edu*

*“Introduction to the Digital Humanities” RCC Workshop*

*Summer 2016*

**Handout for “Tokenization and Transformation: Turning Texts into Databases”**

1. **Whence the Text? (aka “Know thy Source Data”)**
2. **Regular Expressions: Search and Replace**
   1. Always work with a **copy** of your source file
   2. Working in Word
   3. “Wildcards” and Text Editors: Full regular expression capabilities
   4. Preparing the text/data for export
      1. Markup and layout strategies (see also section IV below)
      2. Flat text – save as **UTF-8 (Unicode)**
      3. File formats: Tab-delimited, CSV, JSON (many options)
3. **Word 🡪 Flat Text File 🡪 Excel**
   1. Export to plain text (≠OLE)
   2. Excel Strategies:
      1. Importing
      2. Arranging the data
      3. Generating probabilities and statistics
      4. Visualizing the data (D3: <https://github.com/mbostock/d3/wiki/Gallery>)
4. **Natural Language Processing Strategies, Part-Of-Speech, Treeviews and Text Mining**
   1. Auto-tagging Parts of Speech: Tokenization
      1. Sinica Tagged Corpus Treeview: Intro <http://turing.iis.sinica.edu.tw/treesearch/>
      2. Treeview Example: <http://turing.iis.sinica.edu.tw/treesearch/treeview.exe?sTID=42486>
      3. Tagged Corpus Example:

<http://app.sinica.edu.tw/cgi-bin/kiwi/dkiwi/kiwi.sh?ukey=79872003&qtype=11&tid=0&swTag=1>

* 1. Examples of Tagging and Markup strategies (for Chinese corpora)
     1. MARKUS <http://dh.chinese-empires.eu/beta/>
     2. China Biographical Database Project (CBDB): <http://isites.harvard.edu/icb/icb.do?keyword=k16229&tabgroupid=icb.tabgroup144476>
  2. Creating custom markup procedures
  3. Text-Mining: Upscaling and Downscaling
     1. Bookworm <http://bookworm.culturomics.org/>

Hathi Trust Bookworm Viewer <http://bookworm.htrc.illinois.edu/>

* + 1. Google Books : n-gram Viewer <https://books.google.com/ngrams>
  1. Digital Philology: Philologic4   
     For example: <http://artflsrv01.uchicago.edu/philologic4/shakespeare_demo/>