Jeffrey Tharsen tharsen@uchicago.edu

“Introduction to the Digital Humanities” RCC Workshop

Summer 2017

***Sources:***[*http://tharsen.net/DH\_sources.zip*](http://tharsen.net/DH_sources.zip)

**I. Resources & Tools (See also DDI Life Cycle & Digital Methods Syllabus handouts)**

**A. Building a Corpus / Dataset**

**Existing Corpora**: HathiTrust (22 million volumes available via the Data Capsule), Google Books, many others online:

e.g. ECCO for 18th century literature, gutenberg.org, wikisource, DPLA, COCA/COHA (BYU POS-tagged corpora)

**OCR**: ABBYY (VRC, RCC Walk-in Lab [Reg 216], or purchase), Tesseract (open source), Adobe Acrobat Pro ($$), others

**Cleanup**: Find & Replace (MS Word / regular expressions), OpenRefine

(“building commonplaces” / “creating critical text(s)”)

**B. Data Management + Text Markup (See also “Texts Into Data” handout)**

**“Literacy”** 🡪 ***Iteracy***  (aka “computers are stupid, but they can be trained and/or ‘commanded’ ”) ☺

**Data review**: Excel, LibreOffice Calc, Palladio, RStudio etc.

**Data file types**: CSV, JSON, XML, tab-delimited (TSV), fixed-width

**Databases**: MySQL/SQLite/MariaDB, PostgreSQL, OCHRE (atomized database), NoSQL (flat files), others

(how to select & optimize for your uses)

**Text Auto-Markup**: Stanford NLP (NER, coreferences, POS tagging, tokenization), Samtla, MARKUS (for Chinese), TEI tagging

**Text Visualization and Topic Modeling**: Voyant Tools, MALLET (Topic Modeling), “word clouds”, TAPoR

**Archives of Images**: British Museum, British Library, National Palace Museum, Rumsey Map Collection, many others

**C. Maps and Mapmaking**

**Custom Maps**: ArcGIS (+ESRI Javascript API) / QGIS, OpenStreetMap (+Leaflet), Google Maps/Google Earth (+API),

OMEKA+OpenLayers (Drupal 7/8); animations & layering: Raphael, Canvas objects (Javascript)

**Geocoding** (batch geocoding) : doogal.co.uk/BatchGeocoding.php , batchgeo.com , others

**D. Website Construction**

**Website building**: WordPress (use UChicago Voices “UChicago Unit Website Template” = Divi Theme), Drupal, Omeka

**Connecting** to the server: Mac samba (smb://) mount, PC map network drive; SFTP transfer (FileZilla/FTP client); scp command

**Hosting** options: IT Services, Hum. Comp, Midway (from /home or /project), GoDaddy, many others.

**Sandbox (IDE)**: Midway (via the RCC), on your laptop/desktop: Python Notebook, RStudio, set up an Apache web server, etc.)

**II. Developing Custom Algorithms/Toolkits/Platforms**

**A. Writing Code**

**Programming languages**: Python (Cython), R, Javascript, Perl, PHP, HTML/CSS, Go; C, Java (full application development)

**Programming editors**: TextWrangler (Mac), Notepad++ (Windows), Spyder (Python IDE), Sublime ($), many others

* **Machine learning** strategies (“training the algorithm”; unsupervised —> supervised)
* **Advanced custom** **algorithms** for search, retrieval and analysis (—> scikit learn, RegEx, many others)

**B. Data Visualizations**

**Visualizations**: Tableau, Palladio, D3 (Javascript), gephi / NodeXL (network visualizations); Jupyter Notebook (Python), RStudio