Jeffrey Tharsen tharsen@uchicago.edu

“Introduction to the [Technical Side of the] Digital Humanities” RCC Workshop

Autumn 2021

***Sources:***[*https://github.com/rcc-uchicago/DH\_intro\_workshop*](https://github.com/rcc-uchicago/DH_intro_workshop)

**I. Resources & Tools (See also DDI Life Cycle Model PDF in Sources)**

**A. Building a Corpus / Dataset**

**Existing Corpora**: HathiTrust (16.7 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)

**APIs and Web Scraping:** NYTimes API, Twitter API, JStor API, many others; Social Media scraping, Beautiful Soup (Python)

**OCR**: Tesseract 4.1 (open source), ABBYY (VRC & RCC Walk-in Lab in JRL 216), Google Vision ($), Adobe Acrobat Pro ($)

**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 don’t think, they are “processors”; they can be trained and/or commanded)

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

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

**Databases**: MySQL/SQLite/MariaDB, PostgreSQL, OCHRE (“atomized” graph databases), NoSQL/MongoDB (flat files), others

**NLP**: SpaCy, NLTK, Stanford CoreNLP (tokenization, POS tagging, NER, coreferences); TEI (XML); Word Vectors (word2vec)

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

**Deep Learning Frameworks for NLP & Text Analysis** (sentiment analysis, Q&A, text generation): BERT, GPT (-2, -3, -J)

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

**C. Maps and Mapmaking**

**Custom Maps**: kepler.gl, ArcGIS (ESRI API) / QGIS, CartoDB, OpenStreetMap (+Leaflet), Google Maps/Earth (+API),

OMEKA+OpenLayers (Drupal 7/8); animations & layers: Raphael, Canvas objects (Javascript); 3D: Unity, SketchFab

**Geocoding** (batch geocoding) : batchgeo.com , gis.rcc.uchicago.edu (ESRI/ArcGIS), Google API ($), 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: voices.uchicago.edu (WordPress), Humanities Computing (Drupal); GoDaddy ($), many others.

**Sandbox (IDE)**: Midway2 (RCC HPC) and/or Jupyter Notebook (Python), RStudio; Apache web server

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

* **Machine learning** strategies (“training the algorithm”; unsupervised —> supervised), pattern recognition, clustering
* **Deep learning** strategies : Neural networks (tensorspace.org), transformers (huggingface.co); Quantum computing
* **Advanced custom** **algorithms** for search, retrieval and analysis (RegEx, scikit learn, neural networks)

**B. Data Visualizations**

**Visualizations**: Tableau, gephi (network visualizations), Bokeh (python/R), D3 (Javascript); Jupyter Notebook (Python), RStudio