

# ArchivesZ: Visualizing Archival Collections

## Tackling the Challenges of Data Aggregation



COLLEGE OF  
INFORMATION  
STUDIES

Jeanne Kramer-Smyth ~ <http://www.spellboundblog.com> ~ <http://www.archivesz.org>

### RESEARCH GOALS



Extract and aggregate collection level data from EAD encoded finding aids. Generate cross-collection and cross-repository visualizations.

Photo by Heather Soyka

### BUILDING BLOCKS

Subjects    Linear Feet    Years



What?    How much?    When?  
Data extracted from XML encoded descriptions of archival collections

### NEXT STEPS

- ✦ Collection of additional finding aids
- ✦ Refinement of data extraction logic
- ✦ Support update of finding aids
- ✦ Programmatic removal of stop word tags
- ✦ Institution level configuration files
- ✦ Refinement and completion of Version 2 visualizations

### METHODOLOGY

- ✦ Encoded Archival Description (EAD): international standard for XML encoding descriptions of collections
- ✦ EAD XML files acquired from many repositories
- ✦ Extract, normalize and transfer data to database

**Agricultural colleges – Maryland** – History – Sources  
**Tobacco – Maryland** – History – Sources



Decompose **SUBJECTS** to **TAGS**  
Remove redundant and non-descriptive terms

**Agricultural colleges**  
**Maryland**  
**Tobacco**

### TARGET AUDIENCES

- ✦ Archivists and manuscript curators
- ✦ Researchers, historians and humanities scholars
- ✦ Students

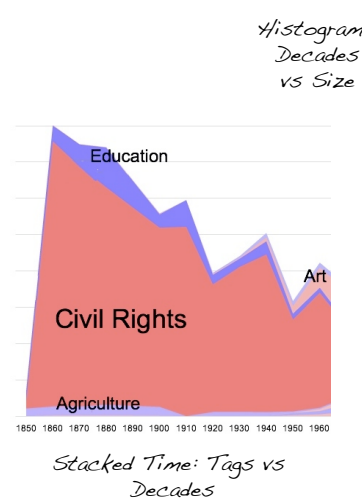
### ArchivesZ Project Credits:

Version 2 Team:    Jeanne Kramer-Smyth  
                              Richard Bovell  
Version 2 Project Funded by NEH Digital Humanities Startup-Grant  
NEH Project Director:    Dr. Jennifer Golbeck  
Version 1 Team:    Jeanne Kramer-Smyth  
                              Tim Anglade  
                              Morimichi Nishigaki

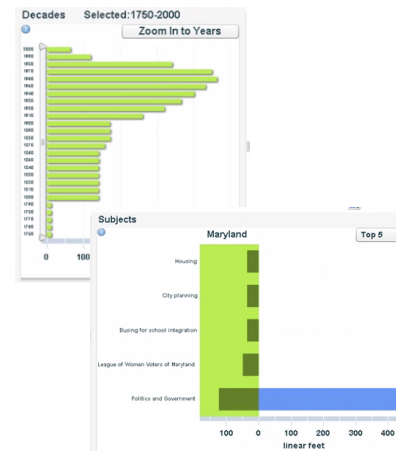
## VISUALIZATIONS OF AGGREGATED DATA

### CHALLENGES

- ✦ Flexibility of EAD encoding guidelines
- ✦ Diversity of descriptive practices and XML encoding standards across institutions
- ✦ Wildly diverse subject terms, both within and among repositories
- ✦ Collection level subjects associated with all records
- ✦ Conversion of all sizes to linear feet



*Histogram: Decades vs Size*



	1910	1920	1930	1940
Civil Rights	Red	Orange	Red	Red
Education	Orange	Yellow	Yellow	Yellow
Art	Green	Orange	Orange	Orange
Agriculture	White	Green	Green	Green

*Heat Map Grid: Tags vs Decades*