Computational Analysis and Visualization of Electronic Records Collections

Visualization and Data Analysis Group
Texas Advanced Computing Center, UT Austin
Research Forum SAA, 2009
Research Motivation

Digital tools for archivists and archives’ users to:

Make sense, explore, appraise, describe, discover, manage, and preserve electronic records collections of varied structures and formats
Challenges

• Large archival data
• Minimal metadata
• What metadata can be extracted from electronic records collections?
• What can be inferred from the structure and the content of electronic records records?
• Abstract representations of collections: visual literacy development
I. Treemaps

• Developed by Ben Shneiderman in the 1990’s
• Adapted to visually scan large collections and to focus on smaller parts
• Collection properties are extracted and stored in a database
  – Structural, descriptive, technical
• Rendered through a treemap visualization application
Directories view
NARA test-bed collections in the Transcontinental Persistent Archives Prototype
Records Group: Records of the Bureau of Census

Smallest partition: one directory
Pink fill: searched keyword (2004) present in the name of the directory
Degrees of yellow-green-brown: from more to less number of files present in each directory / from more to less number of different file extensions present in each directory
II. Paragraph alignment visualization

- Identify related records
  - Belonging to same activities, projects, transactions, events, etc.
  - Content based relationships
<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>C1</td>
<td>D1</td>
<td>D2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C2</td>
<td>D3</td>
<td>D4</td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Compute relationships through bioinformatics-inspired method called “paragraph alignment”
- Unstructured collection of electronic records of different authors, topics, sizes with not much of an a-priori organization
Stories
Discovering context

• Explore and discover relationships between records and their authors
  – Context
    • Evidenced of cooperative writing
    • Work-processes
Conclusions

• Preliminary research
• Usability, display, interoperability and direct adaptation to archival tasks need to be resolved
  – EAD, JHOVE, PREMIS, METS
• Results stored in a database,
  – can be combined in myriad ways and with other tools to make abstractions, synthesis, and new discoveries
• Find useful/new visual metaphors