Eric Walton (BS '09)
2009
Advisor: Dr. G. Dimitoglou

Problem Statement

- 200,000+ Documents
  - Weisberg (1913-2002)
  - FOIA
  - JFK, MLK, RFK
- Physical form
- Digitization Project
Motivation

- Digitized not searchable
- Accessibility limited to on-site viewing
- Web Accessibility
- Physical and Digital
  - not searchable

Problem Statement Cont.

- Metadata?
Related Work

• ARMARIUS
  – Scanned Ancient Documents (JPEG, TIFF, PNG etc…)
    • User Annotation
      • Time Consuming
      • Not Automated

• Google-OCRopus
  – OCR (Optical Character Recognition)
  – Recognizes text and handwriting in PDFs

Challenge

• Metadata
• Ingest and structure
  – Metadata file inadequate
    • Not automatable
  – “Extract” Metadata

<table>
<thead>
<tr>
<th>Remove</th>
<th>File Name</th>
<th>Keyword 1</th>
<th>Keyword 2</th>
<th>Keyword 3</th>
<th>Path</th>
<th>File Type</th>
<th>File Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1472</td>
<td>02-01.pdf</td>
<td>Civil Actions</td>
<td>FOIA Files CA 75-1996/CA</td>
<td>CA 75-1966 FBI Letters 9-12-78 Referrals</td>
<td>/Civil Actions/FOIA Files CA 75-1996/CA 75-1966 FBI Letters 9-12-78 Referrals/02-01.pdf</td>
<td>PDF</td>
<td>1.5mb</td>
</tr>
</tbody>
</table>
Solution: 3-tier Architecture

Pre-Processing Stages
- Parse
- Ingest

Storage Tier
- Disk Repository
- Metadata

Logic Tier
- SRMMS
- Coordination of streams and results output

Presentation Tier
- Submission and return of results from search queries

JFK.hood.edu

The Harold Weissberg Archive, Digital Collection

The Collection
- The Weissberg Archive, sponsored by the Harold Weissberg family, contains over 400,000 documents from various eras of political and cultural history.

Highlights
- Includes the archives of various political figures and organizations.
- Offers digital access to documents that are difficult to find in other physical collections.

Holdings
- Over 400,000 pages of documents from various eras.
- Includes newspapers, political speeches, and personal correspondence.

Sign up for updates!

A Great Place to be Smart