Indexing LANL's Historical Collections in Less Than 1,000 Years

Julie Maze

June 1, 2022

LA-UR-22-24853
Thousands of years to catalog/index

National Security Research Center

Author
Dates
Keywords
Abstract
Orgs/entities
Requirements
- Auto-metadata extraction
- Aid self-guided search and discovery

Key capabilities
- Recognize text
- Extract metadata
- Natural language processing
- Present relevant result
  - search terms
  - relational concepts
6 month pilot

- connected 4 repositories
- tested permission persistence
- tested formats and fonts
- created small taxonomy
Cataloging / Indexing

--> Currently using **fully manual** process
--> Process 10-30 minutes per doc
--> 1.5 FTEs
--> Total number of **digitized files** growing rapidly
--> 14M documents **not yet** digitized

<table>
<thead>
<tr>
<th>Current fully manual process</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>digital</td>
<td>physical</td>
</tr>
<tr>
<td>Quantity</td>
<td>2.4 M</td>
<td>14 M</td>
</tr>
<tr>
<td>Rate/month 1.5 FTEs</td>
<td>486</td>
<td>486</td>
</tr>
<tr>
<td>Years to complete</td>
<td>412</td>
<td>2,400</td>
</tr>
</tbody>
</table>
Cataloging / Indexing

Currently using fully manual process
Process 10-30 minutes per doc
1.5 FTEs
Total number of digitized files growing rapidly
14M documents not yet digitized

<table>
<thead>
<tr>
<th>Current fully manual process</th>
<th>digital</th>
<th>physical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantity</td>
<td>2.4 M</td>
<td>14 M</td>
</tr>
<tr>
<td>Rate/month 1.5 FTEs</td>
<td>486</td>
<td>486</td>
</tr>
<tr>
<td>Years to complete</td>
<td>412</td>
<td>2,400</td>
</tr>
</tbody>
</table>

If we don’t change it will take 2,800+ years to catalog
History is the foundation for the future

Divider: last nuclear test 1992

Knowledge transfer

Current geopolitics

Los Alamos National Laboratory
## What it takes

<table>
<thead>
<tr>
<th>Task</th>
<th>LANL Config</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify repositories</td>
<td>Archive, Shared Drives, Sharepoint Sites</td>
<td>n/a</td>
</tr>
<tr>
<td>Build ontologies</td>
<td>3 FTEs</td>
<td>$500K/yr</td>
</tr>
<tr>
<td>Software</td>
<td>2 dev, 1 enterprise license</td>
<td>$2.5M/yr</td>
</tr>
<tr>
<td>Infrastructure</td>
<td>1PB encrypted storage 0.6FTE</td>
<td>$500K equip $130K FTEs</td>
</tr>
<tr>
<td>Connect repositories</td>
<td>1 FTE</td>
<td>$300K</td>
</tr>
<tr>
<td>QA Testing</td>
<td>TBD</td>
<td>TBD</td>
</tr>
</tbody>
</table>
Can you afford thousands of years?