Decision Points: Creating Non-MARC Genealogy Databases With VuFind

John Houser - October 13, 2014
Historical Society of Pennsylvania
Historical Society of Pennsylvania

• One of the oldest historical societies in the United States

• 600,000 printed items

• Archival collection of 21 million manuscript and graphic items

  • Particularly strong in 17th-, 18th-, and 19th-century

• Active education, digitization, and digital history programs
HSP Web-Presence

- Multiple Drupal websites

- Discovery system based on VuFind

- 206,000 MARC and EAD records

- Digital Library based on Collective Access

  - 13,000 objects
  - 74,000 images

- Many of our records have been digitized by vendors such as Adam Matthew Digital, Ancestry.com, FamilySearch, and NewsBank
We are embarking on a multistep program to fundamentally change how we serve our members online.
WHAT’S THE POINT?

To make HSP’s online resources **findable**, **accessible** and **useful** for non-profit scholarly or personal use from **outside** the building.
Need a better document viewer!

**IMAGE RESOLUTION**

- Current image resolution of many of our images in the Digital Library is **too low** for viewing online
- Switch to **proportional, higher resolution** images
- Make high-resolution images **free** for non-profit use
- Enhance the document viewer to allow **zooming** in at high resolution
Showing 1 - 20 of 4,323 for search: 'civil war', query time: 0.37s

   Published 1861
   “...United States History Civil War, 1861-1865 Causes...”

   Call Number: E458.1.B3x
   Located: Closed Stacks (call slip needed)
   Book Available
Proxy server access to members-only resources such as FamilySearch restricted content and HSP Genealogy
Launch date:

January 2015

February
Pre-release
Search: rush

**Suggested Names**... within your search.

<table>
<thead>
<tr>
<th>Name</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>rush</td>
<td>123</td>
</tr>
<tr>
<td>mrs</td>
<td>10</td>
</tr>
<tr>
<td>more...</td>
<td></td>
</tr>
<tr>
<td>benjamin</td>
<td>11</td>
</tr>
<tr>
<td>william</td>
<td>10</td>
</tr>
<tr>
<td>joseph</td>
<td>10</td>
</tr>
<tr>
<td>jacob</td>
<td>6</td>
</tr>
</tbody>
</table>

**Year**

![Year chart]

Showing 1 - 40 of 123 for search: 'rush', query time: 0.09s

- Select Page | with selected: Email Export Print Save Add to Book Bag

1. **Rush**
   - Year: 1939
   - In database: Genealogical Scrapbooks & Research Folders (1607-1958)
   - "...Rush..."

2. **Rush**
   - In database: Genealogical Scrapbooks & Research Folders

**Sort** Relevance

**Database**

- OHBC (1920-1980) 48
- GS RF (1607-1958) 37
- PRWBM (1775-1783) 34
- HMSP (1883-1889) 3
- PPO (1898-1901) 1

**Narrow Search**

- Year
  - From: [ ]
  - To: [ ]
  - Set
• **Small** databases (first is less than 3,000 records)

• Relatively **small number of fields**

• Not a lot of text

• Often **transcriptions** of handwritten registers or records

• First database: **Philadelphia Home for Infants**

• At least **ten more** coming in the next few years

• **Static** data
DATABASE LOADING

- Staff enter data into Excel spreadsheets
- Perl script to convert to XML
- XML loaded directly into Solr
- **No MARC** loading or index scripts used
Mayo, Mary Elizabeth

1870

30 Jun 1873

18 Jun 1870

Mayo, Jonah

true

210 Willow St.

Taken from the Home by the father on Jan. 23, 1874

Philadelphia (Pa.)

Orphanages--Registers

Registers of births, etc

Adoption--Pennsylvania--19th and 20th centuries--Local sources

Child welfare--Pennsylvania--19th and 20th centuries--Local sources

Orphanages--Pennsylvania--19th and 20th centuries--Local sources
We don’t need a backend database

- The data is static
- Edits can be made to the XML file from which Solr is loaded
- The fields are small (everything can be indexed)

Not as static as I would like
Databases will share a single VuFind instance

- Core fields are similar in each database and can be shared
- Additional indexes can be added for unique fields
- Ease of maintenance and upgrades
Databases will be organized into collections

- VuFind supports hierarchical collections and collection browsing
- Collections may be searched together or separately
Field structure will utilize Solr dynamic fields

- Very little Solr configuration required
- Ease of maintenance and upgrades
Each database will get its own record driver

- Separate record drivers make it easy to create custom templates and displays for each database, even within a shared VuFind instance
Lessons Learned

• Many VuFind functions, and not exclusively the data retrieval functions in the **SolrDefault** and **SolrMarc** record drivers and their associated templates, assume MARC or Dublin Core-like bibliographic fields.

• Think carefully before dumping the **Title** and **Author** fields!

• May be helpful to create record driver for collection records.
REFERENCES

• Loading records into Solr
  https://wiki.apache.org/solr/UpdateXmlMessages

• Hierarchies and collections in VuFind
  https://vufind.org/wiki/hierarchies_and_collections?s[]=collection

• Solr dynamic fields
  https://vufind.org/wiki/adding_facets?s[]=dynamic
  https://vufind.org/wiki/index_schema#dynamic_field_suffixes
9. Benjamin Chew [electronic resource]
Record Source: Digital Records
Record Number: 978
Link to Digital Library record
Electronic

10. Benjamin Chew journal 1758
by Chew, Benjamin 1722-1810
Record Source: Archival Materials
Call Number: Am.043
Located: Closed Stacks (call slip needed)
Manuscript Available

11. Chew, Benjamin, Esquire
Record Source: Names Database
Database: Genealogical Scrapbooks & Research Folders (1607-1958)