Decision Points:
Creating Non-MARC Genealogy Databases
With VuFind

John Houser - October 13, 2014
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
  - 286,000 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
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.
Civil War: its causes, its consequences, its crimes and its compromises / by Beecher, Henry Ward, 1813-1887.
Published 1861
“...United States History Civil War, 1861-1865 Causes...”

Call Number: F458.1.B3x
Located: Closed Stacks (call slip needed)

Book Available
LOG IN SCREEN

Proxy server access to members-only resources such as FamilySearch restricted content and HSP Genealogy
Launch date:

January 2015
Search Results

Suggested Names... within your search.
Pollock, Teresa (3)  Smith, William (3)  Abley, Harry (2)
Allaways, Maggie (2)  Anderson, Harry (2)  Bailey, Zillah May (2)
more ...

Birth Year

Showing 1 - 40 of 2,882 for search: ", query time: 0.09s

Sort  Relevance

Select Page  Add to Book Bag
HSP GENEALOGY

- **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
<doc>
  <field name="id">PHI-52-0004</field>
  <field name="recordtype">Phi</field>
  <field name="hierarchy_top_id">1</field>
  <field name="hierarchy_top_title">HSP Genealogy</field>
  <field name="hierarchy_top_id">2</field>
  <field name="hierarchy_top_title">Philadelphia Home for Infants</field>
  <field name="hierarchy_parent_id">2</field>
  <field name="hierarchy_parent_title">Philadelphia Home for Infants</field>
  <field name="source_txtF">Philadelphia Home for Infants 1873–1924, Volume 52</field>
  <field name="volume_number_txt">52</field>
  <field name="record_number_txt">4</field>
  <field name="primary_name_str">Mayo, Mary Elizabeth</field>
  <field name="primary_name_txt">Mayo, Mary Elizabeth</field>
  <field name="birth_year_str">1870</field>
  <field name="admitted_date_str">30 Jun 1873</field>
  <field name="admitted_date">1873-06-30T00:00:00Z</field>
  <field name="death_date_str"></field>
  <field name="birth_date_str">18 Jun 1870</field>
  <field name="birth_date">1870-06-18T00:00:00Z</field>
  <field name="father_txt">Mayo, Jonah</field>
  <field name="mother_deceased_str">true</field>
  <field name="street_address_txt_mv">210 Willow St.</field>
  <field name="comment_txt_mv">Taken from the Home by the father on Jan. 23, 1874</field>
  <field name="region_txtF">Philadelphia (Pa.)</field>
  <field name="genre_txtF">Orphanages--Registers</field>
  <field name="subject_txtF">Registers of births, etc</field>
  <field name="subject_txtF">Adoption--Pennsylvania--19th and 20th centuries--Local sources</field>
  <field name="subject_txtF">Child welfare--Pennsylvania--19th and 20th centuries--Local sources</field>
  <field name="subject_txtF">Orphanages--Pennsylvania--19th and 20th centuries--Local sources</field>
</doc>
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)
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
DECISION 4

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