Discovery Turned Inside Out

Using schema.org and Google Site Search with Library Digital Collections

Will Sexton
Digital Collections Coordinator

Sean Aery
Digital Projects Developer

CNI Spring 2013 Membership Meeting
Apr 5, 2013
San Antonio, TX
Find us at ...

library.duke.edu/digitalcollections/
library.duke.edu/rubenstein/findingaids/
Goal 2: Provide Digital Content, Tools & Services

"Accelerate digitization of unique Libraries materials and increase access to digital scholarly content in all forms ...."
## Three challenges

<table>
<thead>
<tr>
<th>Organization</th>
<th>Content</th>
<th>Discovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>How does the library select materials for digitization, set objectives, and support the processes?</td>
<td>How do we model facsimiles for a wide array of physical formats, and maintain fidelity to the originals?</td>
<td>How do we help researchers to find and use our collections?</td>
</tr>
</tbody>
</table>
Papyri
<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>He's a Rag Picker</th>
<th>Dublin Core</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Date</strong></td>
<td>1914</td>
<td>Sheet music</td>
</tr>
<tr>
<td><strong>Composer</strong></td>
<td>Berlin, Irving, 1988</td>
<td>Refined Dublin Core</td>
</tr>
<tr>
<td><strong>Illustrator</strong></td>
<td>John Frew</td>
<td></td>
</tr>
<tr>
<td><strong>Lithographer</strong></td>
<td>E.A. Stege</td>
<td></td>
</tr>
</tbody>
</table>
Company
Webcor by Webster-Chicago

Product
Tape Recorder

Magazine ads
Then even the ashes they left behind them, left even the flames and soiled at days they digged as they continue on and become new live fire, and plentifully beautiful clothing to the likes of them when I cannot say a word of the young men.

For what has become of the women? I am positive though they are still alive and well somewhere.

Because this grass tells me there is no such thing as death, preceded all promised, or if even there was, it was at the beginning.

And ceased the moment the first line thing began.

And that nothing but all goes onward and onward...
Dry collodion negatives
<table>
<thead>
<tr>
<th>Track</th>
<th>Image</th>
<th>Title</th>
<th>Description</th>
</tr>
</thead>
</table>

Television ads
<table>
<thead>
<tr>
<th>ID</th>
<th>Interviewee</th>
<th>Birthplace</th>
<th>Interview Location</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>btvct02013</td>
<td>Hendricks, Lola Haynes</td>
<td>Birmingham (Ala.) 1932-12-19</td>
<td>Birmingham (Ala.) 1994-06-22</td>
<td>Civil Service EEOC Investigator</td>
</tr>
<tr>
<td>btvct02020</td>
<td>Archie, Johnnie W.</td>
<td>Lomison (Ala.) 1909-11-10</td>
<td>Birmingham (Ala.) 1994-06-17</td>
<td>Agriculture Manufacturing Construction</td>
</tr>
<tr>
<td>btvct02021</td>
<td>Clarke, Juanita Waiters</td>
<td>Forkland (Ala.) 1923-09-29</td>
<td>Birmingham (Ala.) 1994-06-29</td>
<td>Educator</td>
</tr>
<tr>
<td>btvct02022</td>
<td>Davis, Blanche</td>
<td>Montgomery (Ala.) 1900-07-03</td>
<td>Birmingham (Ala.) 1994-06-28</td>
<td>Domestic Worker</td>
</tr>
<tr>
<td>btvct02025</td>
<td>Brown, George</td>
<td>Dallas County (Ala.) 1909-02-11</td>
<td>Birmingham (Ala.) 1994-06-23</td>
<td>Manufacturing Steel Worker</td>
</tr>
</tbody>
</table>
"This principle holds that [the] significance of archival materials is heavily dependent on the context of their creation, and that the arrangement and description of these materials should be directly related to their original purpose and function."

Steve Hensen
from the SAA's "Glossary of Archival and Records Terminology"
Playing Game on Boat ("Are You There, Bill?")

Original Negative Label
Label text derived from Gamble's handwritten notes.

328 1879  Boat & Honolulu  Are You There, Bill?
How to model this?
<table>
<thead>
<tr>
<th>Newspapers</th>
<th>Early Manuscripts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duke Chronicle, 1959-1970</td>
<td>From a range of traditions</td>
</tr>
<tr>
<td>See LoC's &quot;Chronicling America&quot;</td>
<td>Close study, annotation, and tradition-specific metadata</td>
</tr>
<tr>
<td></td>
<td>Where is the expertise?</td>
</tr>
</tbody>
</table>
The discovery conundrum
Our DIY DC platform.

Supports:

• **Digital collections** (approx. 40 collections, > 70k items)
• **Archival finding aids** (approx. 2000)
• **Finding databases** (5 collections, > 115k records)
• **People pages** (1 item, several more in progress)
The Tripod2 Stack

**DIGITAL COLLECTIONS**
- Discovery
- Access

**FINDING AIDS**
- Access
- Discovery

**PEOPLE PAGES**
- Access
- Discovery

**FINDING DATABASES**
- Access
- Discovery

**UI Layer**
- HTML, CSS, Javascript

**Tripod2 App**
- Search Results
- Search Forms
- Tabbed Facets
- Items
- Portals
- Content Pages
- Finding Aid
- Finding Aid Search Results
- Person / Timeline (e.g., Doris Duke)

**Middleware**
- Django
- Python

**Data Layer**
- SOLR
- SOLR
- METS (XML)
- EAD (XML)
- EAC-CPF (XML)
- METS (XML)
The Tripod2 Stack (cont'd)

**Most Complex Part:**
- HTML, CSS, Javascript

**Used in 12% of Visits:**
- Django

**Data Layer:**
- SOLR
- METS (XML)

**Finding Aids:**
- Access
- Discovery

**Digital Collections:**
- Discovery
- Access
Plan B: Endeca

DIGITAL COLLECTIONS

<table>
<thead>
<tr>
<th>Discovery</th>
<th>Access</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HTML, CSS, Javascript</td>
</tr>
<tr>
<td>UI Layer</td>
<td></td>
</tr>
<tr>
<td>Tripod2 App</td>
<td></td>
</tr>
<tr>
<td>Middleware</td>
<td></td>
</tr>
<tr>
<td>Data Layer</td>
<td></td>
</tr>
<tr>
<td>ENDECA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Item Pages</td>
</tr>
<tr>
<td></td>
<td>Portals</td>
</tr>
<tr>
<td></td>
<td>Context Pages</td>
</tr>
<tr>
<td></td>
<td>Django</td>
</tr>
<tr>
<td></td>
<td>Python</td>
</tr>
<tr>
<td></td>
<td>METS (XML)</td>
</tr>
</tbody>
</table>
Integration

Facets

Thumbnails
What's Plan C?
Plan C: Google?
### Our Google experiment

<table>
<thead>
<tr>
<th>schema.org</th>
<th>sitemaps</th>
<th>Site Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>embed structured data in HTML</td>
<td>targeted indexing</td>
<td>develop around Google APIs</td>
</tr>
<tr>
<td>&quot;things, not strings&quot;</td>
<td>SEO</td>
<td>use Google algorithm, features</td>
</tr>
<tr>
<td>rich snippets</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3 numbers
1 problem

80% - amount of time (estimated) spent by Will & Sean on Tripod2 that involves discovery functionality

12% - number of visits to digital collections site that involve the use of the native search functionality

70,000 - approximate number of items in digital collections

Problem - How can we extend the impact of our work?
External Discovery
Why it Matters
Unless it is particularly significant or central, an institutional resource is unlikely to have strong gravitational attraction. For this reason search engine optimization, syndication of metadata to relevant hub sites, selective adding of links to Wikipedia, and other approaches are becoming of more interest.

Lorcan Dempsey
VP, OCLC Research & Chief Strategist


Libraries will have to more actively promote the broad discoverability of institutional resources.

Lorcan Dempsey
VP, OCLC Research & Chief Strategist


Outside in: Consolidation of licensed?

Move of bought to licensed/electronic?

GBS?  

In many collections

Inside out:
reputation, institutional assets, institutional record, distinctive, ...

Individual destinations? Academic SEO.

Digital Collections Traffic Sources

Direct: 25.5%

Visited via...

- Library catalog
- Library website
- Finding aids
- Subject guides
- Typing URL directly
- Browser bookmark
- Non-Web links (emails, apps, etc.)
Digital Collections Traffic Sources

Referral: 43.9%

Visited via...
- Wikipedia
- Facebook
- Various websites

Social referrals: 5.0% of total
Digital Collections Traffic Sources

Search: 30.6%

Visited via...

- Google
- Bing
- Yahoo
- AOL
- etc.
Source of Search Engine Traffic

Digital Collections 2011-12
is important. How do we optimize the representation of our materials there?
Which pages & images to index?

sitemaps.org  +  Google Webmaster Tools
Man reading newspaper at table in library with "Reading For Fun" sign on shelf behind.

Creator: Gedney, William Gale, 1932-1989

Date: ca 1955

William Gedney Photographs & Writings
Structured Data in HTML

It's hot. And it's getting easier to do.
"The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation."

Tim Berners-Lee
Inventor of the WWW
W3C Director

Structured Data Syntax: Getting Easier

Expressing linked data in HTML

RDF  Micro-formats  RDFa  HTML5 Microdata  RDFa Lite

1999 → 2013
Killer Apps?
KnowledgeGraph" Things, Not Strings"
French Toast

southernfood.about.com/.../frenchtoastrecipe...
Rating: 3.5 - 45 reviews
This is a French toast recipe with eggs and bread. Basic recipe for French toast with eggs and butter.
Ingredients: eggs, sugar, milk, butter, maple syrup

French Toast

www.foodnetwork.com/.../Quick and Easy
Rating: 5 - 555 reviews - 34 mins
Get this all-star, easy-to-follow Food Network French Toast recipe from Alton Brown.

French Toast

www.simplyrecipes.com/recipes/french_toast/
20 mins
Thick slices of bread, soaked in beaten eggs with milk and cinnamon, toasted in a frying pan, and served with butter and maple syrup.
Ingredients: eggs, milk, cinnamon, butter, maple syrup, orange, triple sec

Vanilla French Toast Recipe

www.tasteofhome.com/Recipes/Vanilla-French...
21 reviews - 10 mins - 332 cal
We discovered this recipe in Mexico. We couldn’t figure out what made this French toast so delicious until we learned the secret was vanilla - one of Mexico’s ...
Ingredients: eggs, milk, sugar, vanilla extract, maple syrup

Fluffy French Toast

allrecipes.com/recipe/fluffy-french-toast/
Rating: 4.6 - 1556 reviews - 30 mins - 123 cal
Filter by Property
"Search Tools"

Perfect French Toast Recipe from Elisa DeCristo
www.tastebook.com/recipes/1553012-Perfect-French-Toast
This is one of my favorite recipes. Unfortunately, I've always made it kid friendly without the Cognac. Perfect French Toast Recipe from Elisa DeCristo.

The Best French Toast You'll Ever Make
www.justapinch.com/recipes/the-best-french-
Rating: 5 - 1 vote
6 large eggs. 1 1/2 cups milk. Juice of 1 medium orange, (about 1/4 cup). 2 tablespoons pure vanilla extract. 2 tablespoons cognac, (optional). 1 tablespoon ...
Ingredients: cognac, eggs, milk, vanilla extract, sugar, lemon, cinnamon ...

Perfect French Toast Recipe from Tracy and Andre
www.tastebook.com/recipes/1308023-Perfect-French-Toast
Tried and true! Perfect French Toast Recipe from Tracy and Andre.
Open Graph Social Object Protocol (uses RDFa)

http://ogp.me

The Open Graph protocol

```html
<meta property="og:title" content="The Story of Coffee and How to Make It" />
<meta property="og:description" content="Collection: Emergence of Advertising in America. origins, Abyssinia, processes, etc. Beautiful illustrations. Images of each page from this item are available. Searchable text is only available for the title page, index, and/or contents pages. Chapter headings: Coffee From Abyssinia - Coffee Comes to Europe - Coffee Picking Time - Washing Coffee Berries - Roasting the Coffee - Packing the Coffee - Science Approves Coffee - How to make Maxwell House Coffee - Back Cover." />
<meta property="og:site_name" content="Duke Digital Collections" />
<meta property="og:type" content="image" />
<meta property="og:image" content="http://library.duke.edu/digitalcollections/media/jpg/eaa/med/CK0062-01.jpg" />
```
Schema.org
A search engine-centric vocabulary
What is Schema.org?

This site provides a collection of schemas, i.e., html tags, that webmasters can use to markup their pages in ways recognized by major search providers. Search engines including Bing, Google, Yahoo! and Yandex rely on this markup to improve the display of search results, making it easier for people to find the right web pages.

Many sites are generated from structured data, which is often stored in databases. When this data is formatted into HTML, it becomes very difficult to recover the original structured data. Many applications, especially search engines, can benefit greatly from direct access to this structured data. On-page markup enables search engines to understand the information on web pages and provide richer search results in order to make it easier for users to find relevant information on the web. Markup can also enable new tools and applications that make use of the structure.

A shared markup vocabulary makes it easier for webmasters to decide on a markup schema and get the maximum benefit for their efforts. So, in the spirit of sitemaps.org, search engines have come together to provide a shared collection of schemas that webmasters can use.
## Thing > CreativeWork
The most generic kind of creative work, including books, movies, photographs, software programs, etc.

<table>
<thead>
<tr>
<th>Property</th>
<th>Expected Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Properties from Thing</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>additionalType</td>
<td>URL</td>
<td>An additional type for the item, typically used for adding more specific types from external vocabularies in microdata syntax. This is a relationship between something and a class that the thing is in. In RDFa syntax, it is better to use the native RDFa syntax – the 'typeof' attribute – for multiple types. Schema.org tools may have only weaker understanding of extra types, in particular those defined externally.</td>
</tr>
<tr>
<td>description</td>
<td>Text</td>
<td>A short description of the item.</td>
</tr>
<tr>
<td>image</td>
<td>URL</td>
<td>URL of an image of the item.</td>
</tr>
<tr>
<td>name</td>
<td>Text</td>
<td>The name of the item.</td>
</tr>
<tr>
<td>url</td>
<td>URL</td>
<td>URL of the item.</td>
</tr>
<tr>
<td><strong>Properties from CreativeWork</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>about</td>
<td>Thing</td>
<td>The subject matter of the content.</td>
</tr>
<tr>
<td>accountablePerson</td>
<td>Person</td>
<td>Specifies the Person that is legally accountable for the CreativeWork.</td>
</tr>
<tr>
<td>aggregateRating</td>
<td>AggregateRating</td>
<td>The overall rating, based on a collection of reviews or ratings, of the item.</td>
</tr>
<tr>
<td>alternativeHeadline</td>
<td>Text</td>
<td>A secondary title of the CreativeWork.</td>
</tr>
<tr>
<td>associatedMedia</td>
<td>MediaObject</td>
<td>The media objects that encode this creative work. This property is a synonym for encodings.</td>
</tr>
<tr>
<td>audience</td>
<td>Audience</td>
<td>The intended audience of the work, i.e. the group for whom the work was created.</td>
</tr>
<tr>
<td>audio</td>
<td>AudioObject</td>
<td>An embedded audio object.</td>
</tr>
<tr>
<td>author</td>
<td>Organization or</td>
<td>The author of this content. Please note that author is special in that HTML 5 provides a special mechanism for</td>
</tr>
<tr>
<td></td>
<td></td>
<td>creating a special 'author' attribute for this purpose. The presence of this attribute causes a special, calculated relevance score to be applied.</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Note:</strong> The 'author' property is considered an 'abstract' property, meaning that it is defined for the creative work, not the creator. This property is intended to be used to link to the creator of the content, not to indicate ownership of the content.</td>
</tr>
</tbody>
</table>
What Makes a Rich Snippet?

- Events
- Music
- Organizations
- People
- Products
- Recipes
- Reviews
- Software Apps
- Videos
- Site Breadcrumbs
Man reading newspaper at table in library with "Reading For Fun" sign on shelf behind.

The 5,000 item collection documents Gedney’s work from the 1950s to 1989. Subjects include photographs of cross country road trips; rural New York; Manhattan; Brooklyn; rural Kentucky; Hippies in San Francisco; composers; gay rallies and demonstrations; St. Joseph’s School for the Deaf; India; England; Ireland; France; and, a large number of nocturnal pictures.

Man reading newspaper at table in library with "Reading For Fun" sign on shelf behind.
Using Google for our Search
Challenges & opportunities
Use the products we expect our end users to use.
Google Custom Search
Already in use for library website search
Search Results

About 227 results (0.16 seconds)

Procter & Gamble Detergents and Soaps (procter_gamble_soaps)
Collection: AdViews. Television commercials created for Procter and Gamble from the D'Arcy Masius Benton & Bowles advertising agency archives held in the...
library.duke.edu/digitalcollections/adviews_procter_gamble_soaps/

at Gimbels! Procter & Gamble's new "Lilt"--the home permanent at Gimbels! Procter & Gamble's new "Lilt"--the home permanent that looks... feels. ... behaves so much like naturally curly hair! Loading. Sizes All Sizes; Info; More ...
library.duke.edu/digitalcollections/adaccess_BH0355/

Prell Shampoo (prell) - AdViews - Duke Libraries
Collection: AdViews. Television commercials created for Procter ...
library.duke.edu/digitalcollections/adviews_prell/

Crest 1950s-1960s (crest_50s_60s) - AdViews - Duke Libraries
Television commercials created for Procter and Gamble from ...library.duke.edu/digitalcollections/adviews_crest_50s_60s/

Digital Collections
Already part of the website search
Good. But not good enough for a true discovery interface. Why not?
Search by Property
Google Site Search

A highly customizable version of Google Custom Search.

APIs:
- CSE Control API v2
- CSE Control API v1
- XML API (most power & flexibility)
Example:

JS API v1
Customized Rich Snippets
A few other goodies to explore using the Google APIs...
Search Tools: Color
Lessons Learned

It's early, but so far...
What We've Learned

- Harder than we thought
- Lots of choices
  - microdata vs. RDFa vs. RDFa Lite
  - schema.org mappings
  - mix use with other vocabularies (e.g., Dublin Core)
  - Google API flavors
- Sparse documentation
What We've Learned

- Indexing is not on-demand
- Images get indexed far slower than pages
- "Big Google" rich snippets mostly elusive
Big Google, Before Schema.org Markup
With VideoGallery Markup: Rich Snippet

Google

duke procter and gamble soap commercials

Web Images Maps Shopping More Search tools

About 3,310,000 results (0.21 seconds)

Procter & Gamble Detergents and Soaps - Duke University Libraries
library.duke.edu/digitalcollections/adviews_procter_gamble_soaps/
20+ items – The digitized items from AdViews are searchable using Duke's ...
1 P&G: Enzyme Bold, 1960s-1970s (dmbb47003)
2 P&G: Cheer Detergent, 1960s-1970s (dmbb47004)

Procter & Gamble - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Procter_%26_Gamble
William Procter, a candlemaker, and James Gamble, a soapmaker, emigrated from .... Procter & Gamble also supports many Spanish-language novellas through advertising on ..... Coat of arms of His Royal Highness The Duke of Edinburgh ...

Procter Gamble Soaps : Free Movies : Download & Streaming ...
archive.org › Moving Image Archive › Television › AdViews
Duke University Libraries. 22 itemsWelcome to Procter Gamble Soaps. AdViews is a digital archive of thousands of vintage television commercials dating from ...
What's Next?
# Measure "Big Google" Impact

## Search Queries

<p>| A        | T | U | V | W | X | Y | Z | AA | AB | AC | AD | AE | AF | AG | AH | AI | AJ | AK | AL | AM |
|----------|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|</p>
<table>
<thead>
<tr>
<th>Page</th>
<th>Lib Google Result</th>
<th>Query List</th>
<th>Impressions</th>
<th>Clicks</th>
<th>CTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>45</td>
<td>Digital Collections</td>
<td>view</td>
<td>view</td>
<td>view</td>
<td>view</td>
</tr>
<tr>
<td>46</td>
<td>Digital Collections homepage</td>
<td>view</td>
<td>view</td>
<td>view</td>
<td>view</td>
</tr>
<tr>
<td>47</td>
<td>Ad Access portal</td>
<td>view</td>
<td>view</td>
<td>view</td>
<td>view</td>
</tr>
<tr>
<td>48</td>
<td>Gamble portal</td>
<td>view</td>
<td>view</td>
<td>view</td>
<td>view</td>
</tr>
<tr>
<td>49</td>
<td>Historic American Sheet Music portal</td>
<td>view</td>
<td>view</td>
<td>view</td>
<td>view</td>
</tr>
<tr>
<td>50</td>
<td>Godfrey portal</td>
<td>view</td>
<td>view</td>
<td>view</td>
<td>view</td>
</tr>
<tr>
<td>51</td>
<td>Photograph - Three Girls in Kitchen</td>
<td>view</td>
<td>view</td>
<td>view</td>
<td>view</td>
</tr>
<tr>
<td>52</td>
<td>Photograph - Child in Tiger Costume</td>
<td>view</td>
<td>view</td>
<td>view</td>
<td>view</td>
</tr>
<tr>
<td>53</td>
<td>Advertising Image - Beauty on Duty</td>
<td>view</td>
<td>view</td>
<td>view</td>
<td>view</td>
</tr>
<tr>
<td>54</td>
<td>Advertising Image, Paginated - Durham Whiffs</td>
<td>view</td>
<td>view</td>
<td>view</td>
<td>view</td>
</tr>
<tr>
<td>55</td>
<td>Sheet Music - Alexander's Ragtime Band</td>
<td>view</td>
<td>view</td>
<td>view</td>
<td>view</td>
</tr>
<tr>
<td>56</td>
<td>Video Album - AdViews</td>
<td>view</td>
<td>view</td>
<td>view</td>
<td>view</td>
</tr>
</tbody>
</table>

## Page Impressions

- **/research/citing/**: 170,000 impressions, 10,000 clicks, 5% CTR, Avg. position: 5.7
- **/digitalcollections/hasm/**: 75,000 impressions, 320 clicks, 0% CTR, Avg. position: 11
Assess Local Impact

- Does this give our users a better discovery experience?
  - Feedback forms, usability testing, web analytics
- Do developers spend less time maintaining the index?
Project Progress to Date

1. Establish baseline metrics
2. Add schema.org markup to pages
3. Index via Google
4. Develop result pgs w/custom Rich Snippets
5. Deploy on site
6. Engage partners
7. Assess, enhance, & refine
Thank You

Will Sexton
will.sexton@duke.edu

Sean Aery
sean.aery@duke.edu