Piles of Stuff:
On Aggregating Digital Collections

Paul Conway
University of Michigan School of Information

2016 Digital Commonwealth Annual Conference
How much data is generated every minute?

https://www.domo.com/blog/2015/08/data-never-sleeps-3-0/
“Organic is nice, but haven’t you got anything digital?”
Transitions: 1996-2006
  - Discovery 2 Delivery + Creation 2 Curation

Significance of new for-profit models

Trends ahead from 2006

Digitization and aggregation

- Brief histories

Three super aggregators

- Europeana
- Digital Public Library of America
- Collex (NINES)

The next wave for aggregation

- Education & Training
- Analytics
Digitization is Image Science.

First Digital Image, 1957
Russell Kirsch, National Bureau of Standards

Boyle & Smith, 1969
First Charge-Coupled Device

Steven Sasson, 1975
First Digital Camera

First Flatbed Scanner, 1978
Ray Kurzweil
Digitization in the Cultural Heritage Sector
... from experiments to projects to programs

RLG Digital Image Access Project (DIAP) - 1993-1995

is mostly here, but the need to index and make images available has only become more pressing thereby. The RLG Digital Image Access Project (DIAP) decided to explore what happens when multiple sources of images reflecting many different indexing traditions are searched together. Not surprisingly, McClung reported that this was the hardest and least conclusive project on which she had ever worked. Her approach has always been to get the brightest people together and give them a well formulated problem and then the solution is arrived at. In this case they

“... McClung reported that this was the hardest and least conclusive project on which she had ever worked.”
Top Digitization Guidelines - 2000s

- National Archives (2004)
- North Carolina (2007)
- Colorado (2008)
Federal Agencies Digitization Guidelines Initiative

DRAFT September 2, 2015

Technical Guidelines for Digitizing Cultural Heritage Materials

http://www.digitizationguidelines.gov/
What is an Aggregator?

- *ag·gre·ga·tor* ˈagrəˌɡādər/ *a website or program that collects related items of content and displays them or links to them.*
  - Open Archives Initiative - Protocol for Metadata Harvesting
    - Based on Dublin Core descriptive metadata framework
  - Resource Description Framework - W3C Semantic Web
    - Mapping diverse local collections to a common scheme
  - Most aggregator services assemble metadata only
    - A distributed model designed for scale
Thematic Research Collections

CORPUS THOMISTICUM
INDEX THOMISTICUS
by Roberto Busa SJ and associates
web edition by Eduardo Bernot and Enrique Alarcón
English version

Search:

concordances terms works options new search
Thematic Research Collections

Περσεύς
PROJECT

An Evolving Digital Library on Ancient Greece and Rome

Check out our preliminary plans for a library of Shakespeare's sources and other Renaissance materials!

Copyright | General Policies of this Site

Notice: Due to the annual maintenance on the electrical system at Tufts, this server will be off-line on either the 24th or 25th of May. We expect to restore service on the 26th of May. We apologize for any inconvenience.
Thematic Research Collections

Welcome to Perseus 4.0, also known as the Perseus Hopper.
Read more on the Perseus version history.
New to Perseus? Click here for a short tutorial.

Perseus Digital Library Updates

- **9/1/2015** Collaborating Courses on Fifth-Century Greek History in Spring 2016?
  Call for Collaboration Gregory Crane Leipzig and Tufts Universities September 1, 2015 This is a preliminary call for comment ...

- **8/7/2015** Open Patrologia Graeca 1.0
  August 8, 2015 http://tinyurl.com/nlvhy9b Comments to munson@dh.uni-leipzig.de Federico Boschetti, CNR, Pisa Gregory Crane, ...

- **7/28/2015** Essays on Digital Classics and Digital Humanities
  Gregory Crane Professor of Classics and Winnick Family Chair of Technology and Entrepreneurship Tufts University Alexander von ...

- **7/28/2015** "Privashing" vs. publishing — in search of an accurate term for a problematic academic tradition
  Gregory Crane Comments to gcrane2008@gmail.com [Draft as of July 28, 2015] Alexander von Humboldt Professor of Digital ...

- **7/20/2015** Revision to "The Big Humanities, National Identity and the Digital Humanities in Germany"
  I have integrated a number of comments and clarifications from my colleagues in Dariah-DE and TextGrid in revising "The …

- **7/20/2015** The Work of Classical Studies in a Digital Age
  Gregory Crane gcrane2008@gmail.com [Draft as of July 20, 2015] Alexander von Humboldt Professor of Digital Humanities Leipzig …

Popular Texts

- Caesar, Gallic War (English, Latin)
- Catullus, Carmina (English, Latin)
- Cicero, In Catilinam I (English, Latin)
- Vergil, Aeneid (English, Latin)
- Herodotus, Histories (English, Greek)
- Homer, Odyssey (English, Greek)
- Plato, Republic (English, Greek)
- Tom Martin. Overview of Classical Greek History from Mycenaean to Alexander (English)

Art and Archaeology

<table>
<thead>
<tr>
<th>Aegina, Temple of Aphaia</th>
<th>Silver obol from Athens</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Barletti Head</td>
<td>Satyr on Attic red figure vase</td>
</tr>
</tbody>
</table>
Aggregation - Digital Library Origins
Thematic Collections and Digital Humanities

The Walt Whitman Archive

Folsom, Ed - The Canary Professor of English at the University of Iowa
Price, Kenneth M. - The Hilegass Professor of American literature at the University of Nebraska, Lincoln
Networled Associate Fellows 1997

The Walt Whitman Archive is an electronic research and teaching tool that sets out to make Whitman's vast work, for the first time, easily and conveniently accessible to scholars, students, and general readers. Whitman, America's most influential poet and one of the four or five most innovative and significant writers in

IATH: http://www.iath.virginia.edu/
Making of America - Bound and Structured
Google Books and HathiTrust Digital Library

Search
Book Search works just like web

Browse books online
If the book is out of copyright, or the

Buy books or borrow from
the library

Learn more fast
We've created reference pages for
Google Books and HathiTrust Digital Library
Key Notes

Digitization and aggregation
  • Brief histories

Three super aggregators
  • Europeana
  • Digital Public Library of America
  • Collex (NINES)

The next wave for aggregation
  • Education & training
  • Analytics
Emerging Lessons from Three Aggregators
Europeana Collections

http://www.europeana.eu/portal/
Europeana Collections - Innovations

+ path breaking standards compliance and development
  - Resource Description Framework (RDF)
+ extraordinary progress on metadata manipulation
+ technical documentation optimized for developers
+ innovation in visualization
+ alliances with computer/info science researchers
# Europeana Publishing Framework

Here you see how the different tiers of the publishing framework work. In this example we use images. For video, audio, text and 3D, please see the Europeana Publishing Framework Appendix.

<table>
<thead>
<tr>
<th>TIER</th>
<th>DIRECT LINK TO OBJECT PROVIDED</th>
<th>RIGHTS STATEMENTS</th>
<th>EUROPEANA API DISPLAYS</th>
<th>BENEFITS (CUMULATIVE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Europeana as a search engine. 'I want people to find my collections’</td>
<td>No</td>
<td>Any from the Europeana licensing Framework</td>
<td>Metadata, no direct link to object</td>
</tr>
<tr>
<td>2</td>
<td>Europeana as a showcase. ‘I want people to find and view my collections on Europeana’</td>
<td>Yes. Minimum of 800+ pixels wide</td>
<td>Any from the Europeana licensing Framework</td>
<td>Metadata plus direct link to object</td>
</tr>
<tr>
<td>3</td>
<td>Europeana as a distribution platform for non-commercial re-use. ‘I want people to find, view and use my collections - but not for commercial gain’</td>
<td>Yes. Minimum of 1,200+ pixels wide recommended</td>
<td>Any from the Europeana licensing framework that allow for some re-use</td>
<td>Metadata plus direct link to object and can filter to show only re-usable objects</td>
</tr>
<tr>
<td>4</td>
<td>Europeana as a free re-use platform. ‘I want people to find, view and use my content - however they want’</td>
<td>Yes. Minimum of 1,200+ pixels wide recommended</td>
<td>Any from the Europeana licensing framework that allow free and open access</td>
<td>Metadata plus direct link to object and can filter to show only re-usable</td>
</tr>
</tbody>
</table>
Digital Public Library of America

http://dp.la/
DPLA -- Innovations

- Hubs and Service Hubs distribute effort and commitment
- Extraordinary documentation for API developers
- Cross connections to K-12 education [public library!]
- Strong commitment to books [HathiTrust/Google]
- Tools for geospatial, temporal, and thematic displays
Collex Search Results
Collex -- Innovations

- Conceived, developed, and lead by scholar-users
- Peer review of collection contributions
  - Selectivity improves overall quality
- Strong commitment to internal analysis tools
  - Juxta for juxtaposition and annotation
  - Commentary
  - Personal collections
- Efforts to foster a publishing environment
  - Exhibits, attempts at open access journals
Aggregating Great Lakes Environmental History
- www.greatlakescollections.org
DPLA Metadata Application Profile (MAP)

Figure 1. DPLA MAP 4.0 Class Diagram

DPLA Where to Start?

- Prospective partners ... test their standards against DPLA’s expectations ...
  - ... hubs are responsible for data quality
  - ... make sure data is as error-free as possible
  - ... elements and properties are consistently implemented
  - ... contextualize your data on a global level
  - ... descriptions useful to an unfamiliar audience
  - ... field tags are internally consistent across sub-collections
Europeana - Focus on Metadata Quality

“... Accessibility, accuracy and consistency of metadata and content are hugely important for the service we want to develop with you, our data partners.”

Every metadata record must have
- dc:title or dc:description
- dc:language (texts)
- dc:subject or dc:type or dc:spatial or dc:coverage
- edm:dataProvider (source institution to aggregator)
- edm:provider (aggregator)
- edm:isShownAt (URL link to item)
- edm:rights (intellectual property)
- persistent identifier

“Hubs” or “Twice Removed”?

```
great-great-grandfather = great-great-grandmother

= great-grandmother

grand-uncle or grand-aunt

first cousin twice removed

first cousin once removed

second cousin once removed

second cousin

second cousin once removed

second cousin twice removed

third cousin once removed

third cousin

third cousin twice removed
```

Content Hubs

Service Hubs
Pass Through to Source/Provider
Key Notes

Digitization and Aggregation
  • Brief histories

Three super aggregators
  • Europeana
  • Digital Public Library of America
  • Collex (NINES)

The next wave for aggregation
  • Education & Training
  • Analytics
Collective action on D2D services, including ...

- Unified, syndicated, and extended discovery services
  - Progress!
- Virtual reference networks
  - Barely attempted
- Aggregated user feedback
  - Not in the frame, yet

Educate and Train for Life Beyond Search

- Anticipate the impact of RDF aggregation on users and use
- Give care to derivative images [and the landing interface]
- Explore the lingering value of “hier-archival” context
- Embrace the full curation lifecycle - D2D & C2C
- Curate cultural heritage organizations
Compete with Data Analytics

- Preserve ethical commitment to privacy and confidentiality

- Take a page from Google and Amazon

- Capture and use data on search, discovery, transactions, feedback to drive the experience of aggregation
Explore historical collections from libraries, museums, and archives across Massachusetts

Map
Use the map to search for materials from locations across the globe.

Collections
Explore curated collections of historic materials grouped by topic, date, location, or format.

Institutions
View materials from libraries, historical societies, museums, and other cultural institutions.

Formats
Photographs, maps, postcards, manuscripts, books, artifacts, and more.
Thank you for your attention!

Paul Conway
Associate Professor
pconway@umich.edu

University of Michigan School of Information
References [1]

- History of Google Books: https://www.google.com/googlebooks/about/history.html
References [2]

References [3]

- Resource Description Framework. [Link](http://www.w3schools.com/webservices/ws_rdf_intro.asp)
- Scholars Portal. [Link](http://www.scholarsportal.info/)