Title
The Digital Archive: The Data Deluge arrives in the Humanities

Permalink
https://escholarship.org/uc/item/2cc978tf

Author
Borgman, Christine L.

Publication Date
2010-05-10
The Digital Archive: The Data Deluge arrives in the Humanities

Christine L. Borgman
Professor & Presidential Chair in Information Studies, UCLA

“Time Will Tell, But Epistemology Won't: In Memory of Richard Rorty”
A Celebration of Richard Rorty's Archive
UC-Irvine, May 14, 2010
UC-Irvine, May 14, 2010
Deluge!!!

Data!!

Historians

Literary Theorists

Philosophers

Classicists

Image courtesy of Ian Foster, Argonne National Labs
I’m finding that something as simple as constructing my maps of related concepts are not easily applied to primary sources in digital libraries. So what use are the digital libraries, if all they do is put digitally unusable information on the web? The digital libraries don’t offer a platform for traditional note taking, much less for larger scale analysis, either quantitative or qualitative.

UCLA doctoral student in history, personal communication, June 8, 2009 (used with permission)

Files retrieved from Richard Rorty's 3.5-inch floppy disks during the processing of his papers.

Richard Rorty (1931-2007) was a pragmatic philosopher, critical theorist, and public intellectual who is commonly described as one of the most important thinkers of his era. Included in this collection are electronic word-processing files, created between 1988 and 2003. The files were retrieved from Rorty's 3.5" floppy disks during processing of his personal papers and converted to PDF format in order to facilitate preservation and access.

Included are letters, many drafts of writings, lecture notes, syllabi, and exams. Also included are bibliographies of his work, a few administrative files, and a couple of documents relating to his children. This collection also contains drafts of writings done by some of his peers and colleagues. Researchers will notice some data loss and corruption has occurred in some of the earliest files.

Access to Rorty's born digital files is provided through UCI space@ the Libraries. Interested researchers should fill out the researcher application and email it to apcoll@ucd.edu or fax it to (949)824-2472. Access may be granted in less than 5 business days.

These digital files are part of the larger collection, the Richard Rorty Papers, which are described in a finding aid in the Online Archive of California. Access to these non-digital materials is provided in the Special Collections and Archives Reading Room at the UC Irvine Libraries.

Richard Rorty donated his papers to the Critical Theory Archive at the UC Irvine Libraries in 2006. Among the boxes of materials were more than 70 floppy disks containing about 1,100 word-processing files. These born-digital files were copied from the disks and saved in pdf format for long-term preservation and access. They are made available exclusively through the virtual reading room in UCI space@ the Libraries.

Copyright

This material is provided for private study, scholarship, or research. Transmission or reproduction of any material protected by copyright beyond that allowed by fair use requires the written permission of the copyright owners. The authors or their heirs retain their copyrights to the material. Contact the University of California, Irvine Libraries, Special Collections and Archives for more information (apcoll@ucd.edu).

Subcollections in this collection

Advanced Search

<table>
<thead>
<tr>
<th>Conjunction</th>
<th>Search type</th>
<th>Search for</th>
</tr>
</thead>
<tbody>
<tr>
<td>AND</td>
<td>Full Text</td>
<td>derrida</td>
</tr>
<tr>
<td>AND</td>
<td>Full Text</td>
<td>1999</td>
</tr>
</tbody>
</table>

Results/page: 10
Sort items by: Relevance
in order: descending

Go
Advanced Search

Conjunction Search type Search for

AND : Full Text : derrida
AND : Full Text : 1999

Results/page 10 25
Sort items by Relevance in order descending

Go

Your query "((derrida AND (1999))" produced 1 result(s).

Search Results for Subcollection: Richard Rorty born digital research files, 1992-2002

Now showing items 1-1 of 1

1

Gumbrecht, Hans Ulrich. "Socrates und Uncle Sam: Wie anders Richard Rorty die Rolle des Philosophen vorbereitet" draft
http://special.lib.uci.edu. email: spoel@uci.edu, 2010-02-17)
Data Created: 1998-10-13T15:47:20PDT

Now showing items 1-1 of 1

1
Gumbrecht, Hans Ulrich. "Socrates und Uncle Sam: Wie anders Richard Rorty die Rolle des Philosophen verkörpert:" draft

Item File(s)

<table>
<thead>
<tr>
<th>Files</th>
<th>Size</th>
<th>Format</th>
<th>View</th>
</tr>
</thead>
<tbody>
<tr>
<td>gumbrechts article on rorty.pdf</td>
<td>48.44kb</td>
<td>PDF</td>
<td>View/Open</td>
</tr>
</tbody>
</table>

The following license files are associated with this item:

- Original License

Item Description

Title: Gumbrecht, Hans Ulrich. "Socrates und Uncle Sam: Wie anders Richard Rorty die Rolle des Philosophen verkörpert:" draft

Author(s): Gumbrecht, Hans Ulrich

Contributor(s): Rorty, Richard

Date Created: 1998-10-13

Permanent Link to This Record: http://hdl.handle.net/10575/119

Type: Archives and Manuscripts

Language: ger

File Format: application/pdf

Access Rights: To access this item, interested researchers should fill out the researcher application (http://www.lib.ucir.edu/libraries/collections/special/using/patron_registration.pdf) and email it to spooll@uci.edu or fax it to (949)824-2472. Access may be granted in less than 3 business days.

Rules for Use of Special Collections and University Archives

- Place all personal property, including briefcases, purses, backpacks, laptop cases, and notebooks in a locker. Coats must be placed in a locker or on the coat rack. Turn your cell phone to vibrate and answer it in the hall, away from the door. Laptop computers, small wallets, notepaper and pencils may be taken into the reading room.

- The reading room is a no-pen zone. Only pencils and laptop computers may be used to take notes.

- Request all items from the reference librarian. Special Collections and Archives materials are not browseable. Request items or photocopying no later than 15 minutes before department closes.

- There is no self-service photocopying of collection materials. The reference librarian determines what may be copied and staff does all copying. Copies are 20¢ each, and patrons must provide a debit card—available in Copy Services—in order for photocopying to be done.

- Handle and read materials only at the reading room tables. Limit handling of items to the minimum necessary for your research and exercise all possible care to prevent damage to materials.

- Do not mark, take notes on or trace on top of library materials. Keep all materials flat on the table; do not place items in your lap or hold them up.

- Please maintain quiet in deference to your fellow researchers.

- The Reading Room is reserved for patrons using Special Collections and Archives materials. The public computer in the Reading Room is reserved for using ANTPAC, OAC, and other online tools supporting Special Collections materials.

- No eating, drinking or gum chewing is permitted in the reading room.

- If you are using archival or manuscript material:
  - Retain existing order and arrangement of all unbound and manuscript materials. If an item appears out of order, notify the reference librarian, but do not make the correction yourself.
  - For permission to publish or to quote from manuscripts and other unpublished materials, please contact the Special Collections and Archives Public Services Librarian (spcoll@uci.edu).

- Please return all materials to the reference librarian before leaving for the day or for an extended period. If you must leave the reading room temporarily, make arrangements with the reference librarian.

- Upon request, submit all personal property for inspection before removing it from the reading room.
How Tweet It Is!: Library Acquires Entire Twitter Archive

April 14th, 2010 by Matt Raymond

[UPDATE: We posted an FAQ on April 28]

Have you ever sent out a "tweet" on the popular Twitter social media service? Congratulations! Your 140 characters or less will now be housed in the Library of Congress.

That's right. Every public tweet, ever, since Twitter's Inception in March 2006, will be archived digitally at the Library of Congress. That's a LOT of tweets, by the way: Twitter processes more than 50 million tweets every day, with the total numbering in the billions.

We thought it fitting to give the initial heads-up to the Twitter community itself via our own feed @librarycongress. (By the way, out of sheer coincidence, the announcement comes on the same day our own number of feed-followers has surpassed 50,000. I love serendipity!)

We will also be putting out a press release later with even more details and quotes. Expect to see an emphasis on the scholarly and research implications of the acquisition. I'm no Ph.D., but it boggles my mind to think what we might be able to learn about ourselves and the world around us from this wealth of data. And I'm certain we'll learn things that none of us now can even possibly conceive.

Just a few examples of important tweets in the past few years include the first-ever tweet from Twitter co-founder Jack Dorsey, President Obama's tweet about winning the 2008 election, a set of two tweets from a photojournalist who was arrested in Egypt and then freed because of a series of events set into motion by his use of Twitter, and two tweets from the Library of Congress. (UPDATE: Here's the press release.)

Twitter plans to make its own announcement today on its blog from "Chirp," the Official Twitter Developer Conference, in San Francisco.

So if you think the Library of Congress is "just books," think of this: The Library has been collecting materials from the web since it began harvesting congressional and presidential campaign websites in 2000. Today we hold more than 167 terabytes of web-based information, including legal blogs, websites of candidates for national office, and websites of Members of Congress.

We also operate the National Digital Information Infrastructure and Preservation Program www.digitalpreservation.gov, which is pursuing a national strategy to collect, preserve and make available significant digital content, especially information that is created in digital form only, for current and future generations.
ArtHistory.viz | Mining 20,000,000+ Images of Art

ArtHistory.viz uses cultural analytics approaches to explore patterns in art, architecture, and visual culture. Techniques include image processing and computer vision, statistical analysis, information visualization, and scientific visualization.

Gu, Jia. Art Objects as Data Points.
Presentation at Caltech Summer Researchers Poster Session. 2008 (pdf)
Unprecedented Data Availability
Bear River – A First Look

Stream gage comings and goings
1958-1988

Annual Mean Daily Discharge

Daily Average Discharge Availability by State, County

Monthly Mean Daily Discharge by Month

Weekly Mean

Heavy flow years and likely nearby stations easy to spot
Blue = BEAR RIVER NEAR COLLINSON, UT and
B R BASIN OUTFLOW ACCT HWY 83 NR CORRINE, UT
Red = BEAR RIVER NEAR CORRINE, UT
Green = BEAR RIVER AT IDAHO-UTAH STATE LINE, UT

Tendency to get two high flows – winter and snowmelt?

Gages added in Cache and Box Elder county
Utah circa 1960 are discontinued by 1988!

Slide courtesy of Catherine van Ingen, Microsoft Research
FLUXNET: A Network of Networks

- 467 towers worldwide
- 967 site-years of sensor data from 253 towers
- ~20 sensor measurements per tower; 20 derived science variables
- 145 ancillary variables

- Original data set assembled and processed in 2007
- 20x larger than previous synthesis dataset
- 75 paper teams with over 200 scientists

http://www.fluxdata.org

Slide courtesy of Catherine van Ingen, Microsoft Research
Data, data everywhere

Information has gone from scarce to superabundant. That brings huge new benefits, says Kenneth Cukier (interviewed here)—but also big headaches

Feb 25th 2010 | From The Economist print edition

When the Sloan Digital Sky Survey started work in 2000, its telescope in New Mexico collected more data in its first few weeks than had been amassed in the entire history of astronomy. Now, a decade later, its archive contains a whopping 140 terabytes of information. A successor, the Large Synoptic Survey Telescope, due to come on stream in Chile in 2016, will acquire that quantity of data every five days.

Astronomical amounts of information can be found closer to Earth too. Wal-Mart, a retail giant, handles more than 1m customer transactions every hour, feeding databases estimated at more than 2.5 petabytes—the equivalent of 167 times the books in America’s Library of Congress (see article for an explanation of how data are quantified). Facebook, a social-networking website, is home to 40 billion photos. And decoding the human genome involves analysing 3 billion base pairs—which took ten years the first time it was done, in 2003, but can now be achieved in one week.
The Knowledge Mashup

It's the world according to YOU. It's about you getting the information you want when you want it. It's about accessing content from open and collaborative sources, then filtering and focusing that content to meet unique documentation, training, and other educational needs. It's personalized, real-time information delivered directly to your computer devices. That's the promise of Web 2.0 moving to Web 3.0—automatically accessing and controlling the best information you deem as relevant to your needs. In your world, you are the one who provides context and meaning in the cloud of overwhelming data and disparate information. You are the existentialist of the Information Age who defines real knowledge amongst the chaos and chatter.

That's the idea behind the Knowledge Mashup. Find the best pieces of information, tag that information, and then structure and sequence it for specific needs. Use a simple object-oriented approach that identifies articles, videos, pictures, and people as independent objects and then give these objects context and navigation by overlaying them with some type of structural paradigm (a list, content map, outline, search filter).

The idea of a knowledge mashup is to bring together disparate but germane resources for a specific topic. The original author of knowledge mashup can tailor these resources and give context to educate, communicate, and impart information. It allows for usage of the best content that is controlled by an author or teacher in providing context and organization. It is a filter on the fire hose on the Internet to direct just the information needed for a specific product, process, instructional aid, or discipline. Among other benefits, it allows the author to be a “guide by the side” of the reader in finding information rather than the all-knowing “sage on the stage.”
Business mashups enable individual efficiency and innovation

Business mashups offer organizations a way to rapidly adapt to changing business needs. In a 2008 survey conducted by IBM, a majority of CEOs rated their organization's ability to manage change 22% lower than their expected need for change. Business mashups can give people the tools needed to adapt to change, develop new insights, and act on new business opportunities.

With IBM® Mashup Center, organizations can unlock and transform diverse sources of information into mashable assets such as feeds and widgets. These assets can then be dynamically assembled into new applications that address daily business challenges. Enabling business users in this way can reduce application backlog, lower development and increase the reuse of existing assets to reach more people cost-effectively.
what is a mashup?

According to the Wikipedia, a mashup 'combines data or functionality from two or more external sources to create a new service. The term mashup implies easy, fast integration, frequently using open APIs and data sources to produce results that were not the original reason for producing the raw source data. An example of a mashup is the use of cartographic data to add location information to real estate data, thereby creating a new and distinct Web services that was not originally provided by either source'. Practically speaking, Enterprise Mashups combine and remix data from databases, spreadsheets, websites, Web Services, RSS/Atom feeds, and unstructured sources that deliver actionable information for better decision-making.

why is there a need for a 'mashup language'?

An open enterprise mashup markup language can greatly improve the portability of mashup designs and the interoperability of mashup solutions, thereby reducing the risk and cost to organizations that are adopting enterprise mashups.

A Domain Specific Language (DSL) designed specifically for mashups can also improve the quality of mashups of well by accommodating the important design characteristics that make mashups easier to create and reuse.

what is the Open Mashup Alliance (OMA)?

The Open Mashup Alliance (OMA) is an organization charted to promote the adoption of mashup solutions in the enterprise through the evolution of enterprise mashup standards like on open enterprise mashup markup language.
Scholarly Information Infrastructure

• Cyberinfrastructure, eScience, eSocial Science, eHumanities, ... eResearch

• Goal: enable new forms of scholarship that are
  • information-intensive
  • data-intensive
  • distributed
  • collaborative
  • multi-disciplinary
• Scale
• Language and communication
• Space and time
• Social networking

What are data?

Categories of data*

- Observational
- Computational
- Experimental
- Records

Objective or subjective?

- Facts
- “alleged evidence” (Buckland, 2006)

*Long-Lived Data, NSF, 2005
Scientific Data

- **Examples**
  - Ecology: weather, ground water, sensor readings, historical record
  - Medicine: x-rays
  - Chemistry: protein structures
  - Astronomy: spectral surveys
  - Biology: specimens
  - Physics: events, objects
  - Documentation: Lab and field notebooks, spreadsheets

- **Sources**
  - Generate own data
  - Acquire from collaborators, other scientists
  - Data repository
Social Scientific Data

• **Examples**
  - Opinion polls
  - Surveys, interviews
  - Mass media
  - Laboratory experiments
  - Field experiments
  - Demographic records
  - Census records
  - Voting records
  - Economic indicators

• **Sources**
  - Generate own data
  - Acquire from other scholars
  - Data repositories: Social Surveys
  - Government records
  - Corporate records

http://www.census.gov/population/cen2000/map02.gif
Humanities and arts data

• Examples
  • Newspapers
  • Photographs
  • Letters
  • Diaries
  • Books, articles
  • Birth, death, marriage records
  • Church records
  • Court records
  • School and college yearbooks
  • Maps...

• Sources
  • Libraries, archives, museums
  • Public records
  • Corporate records, mass media
  • Acquire from other scholars
  • Data repositories: Beazley, Arts & Humanities Data Service (UK)
Enabling Virtual Conversations

Collaboration Centric View

Data Centric View

Slide courtesy of Catherine van Ingen, Microsoft Research
Why openness matters

• Interoperability trumps all
  • Import and export in open formats
  • Mixup and mashup
  • Add value
  • Avoid lock in
• Discoverability of related
  • Documents
  • Data
  • Assorted digital objects
• Usability and reusability
  • For research
  • For learning
Data access comparison*

• **Data-Centric Science**
  – Data usage
    • Instant access
    • Robotic mining
    • Mashup with other sources
    • Multi-variate analyses
  – Tools
    • Open standards
    • Common platforms
    • Mining and visualization
  – Intellectual property
    • Open access license
    • Share and share alike
    • Derivative works encouraged

• **Archives-Centric Humanities**
  – Data usage
    • Delayed access
    • Manual mining
    • Inspect individual archive
    • Deep interpretation
  – Tools
    • Collection-level finding aids
    • Content specific methods
  – Intellectual property
    • Registered scholars
    • Licensed for local use
    • On-site access

*As noted in the discussion, this slide is a deliberately sharp contrast to make a point; the comparisons are, of course, nuanced and complex.
Rights to Use and Reuse

INTELLECTUAL PROPERTY:
Rights & the Modern University:
How to Move Forward into the 21st Century

Who can read?
Who can learn?
Who can speak?

http://interactive.usc.edu/members/sanderson/IPforum.jpg

http://www.loc.gov/exhibits/civilrights/images/civilrights-homeimage-previ.jpg
I’m finding that something as simple as constructing my maps of related concepts are not easily applied to primary sources in digital libraries. So what use are the digital libraries, if all they do is put digitally unusable information on the web? The digital libraries don’t offer a platform for traditional note taking, much less for larger scale analysis, either quantitative or qualitative.

UCLA doctoral student in history, personal communication, June 8, 2009 (used with permission)
Summary

• The data deluge is real

• Data value lays in
  • Scale
  • Aggregation
  • Analytical tools
  • Distributed access
  • Standard platforms

• Data access depends on
  • Open tools and services
  • Intellectual property agreements

Acknowledgements & Thanks

• National Science Foundation
  – CENS: Cooperative Agreement #CCR-0120778, D.L. Estrin, UCLA, PI.
  – CENS Education Infrastructure: #ESI- 0352572, W.A. Sandoval, PI; C.L. Borgman, co-PI.
  – Towards a Virtual Organization for Data Cyberinfrastructure, #OCI-0750529, C.L. Borgman, UCLA, PI; G. Bowker, Santa Clara University, Co-PI; Thomas Finholt, University of Michigan, Co-PI.
  – Monitoring, Modeling & Memory: Dynamics of Data and Knowledge in Scientific Cyberinfrastructures: #0827322, P.N. Edwards, UM, PI; Co-PIs C.L. Borgman, UCLA; G. Bowker, SCU; T. Finholt, UM; S. Jackson, UM; D. Ribes, Georgetown; S.L. Star, SCU)

• Microsoft External Research

Special thanks for critical commentary on an earlier version of this talk
UCLA Department of Information Studies:
Evan Baker, David Fearon, Monica Garcia, Matthew Mayernik, Ellen Mulligan, Katie Shilton, Derek Quezada, Jillian Wallis, Laura Wynholds

And for advice on intellectual property issues for archives from Maureen Whalen, Getty Trust