“Methodology for the Infinite Archive”: Exploring the Implications of Digital Resources and Tools for Literary Scholarship

Lisa Spiro
April 2009
How are “traditional scholars” using digital collections?

(Morris Eaves at the TAMU Digital Textual Studies Symposium, 2006)

- Focus on “second-order” scholarship—not building collections & tools, but using them to produce research
“Methodology for the Infinite Archive”

“The web is an archive that is constantly changing and effectively infinite. What kind of research techniques can historians develop to make use of it?” (Bill Turkel, Digital History Hacks)
Agenda

I. How are scholars using digital resources?
   ◦ Study of American lit scholars (Spiro/Segal)

II. What does it mean—and take—to produce digital scholarship in the humanities?
   ◦ Dissertation remix project to explore implications of the digital for:
     • Collecting information
     • Analyzing information
     • Disseminating research

III. What are some of the technical challenges facing digital scholarship, and what are strategies to address them?
I. How are scholars using digital resources?

http://tinyurl.com/cklu2c
Overview of Study of the Impact of Digital Resources on American Studies

- General survey of Americanists about how they use digital resources, + follow-up interviews
- Citation analysis of works on Whitman, Dickinson and *Uncle Tom’s Cabin* to determine if they cited major digital collections
- Surveys of Dickinson, Whitman & UTC scholars about their (non-)usage of these digital collections, + follow-up interviews with selected scholars
- Research was conducted in Spring 2007; collaboration w/ Jane Segal
Scholars’ Perception of the Impact of Digital Resources

- General survey question: “Do you think that the availability of electronic resources has transformed humanities scholarship?”

- Positive Impacts
  - Makes research faster, more convenient
  - Increased access to a larger range of resources
  - Allows scholars to ascertain the state of scholarship
  - Allows for new approaches to scholarship

- Negative impacts
  - Information overload
  - Encourages laziness
  - Increased pressure to produce
Scholars Use Digital Collections More Than They Cite Them

Scholarly Works Citing Digital Archives, 2000-2008

<table>
<thead>
<tr>
<th>Archive</th>
<th>% Citing</th>
<th># Citing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whitman</td>
<td>21%</td>
<td>65 of 317</td>
</tr>
<tr>
<td>Uncle Tom</td>
<td>10%</td>
<td>8 of 82</td>
</tr>
<tr>
<td>Dickinson</td>
<td>12%</td>
<td>36 of 294</td>
</tr>
</tbody>
</table>

- 58% of the scholars we surveyed said that they frequently *use* digital resources, but only 26% said that they *cite* them frequently.
The Walt Whitman Archive Is Changing Scholarly Practice

Interviewees & survey results suggested that as a result of the WWA:

- Manuscript & textual study of Whitman have grown
  - “Now you can look at different images and books and can chart how a poem has evolved from edition to edition.”

- Scholars can more easily use related materials, e.g. Traubel’s *With Walt Whitman in Camden*
  - “It seems like more Whitman scholars are citing Traubel as a result of its being more readily available through the *Walt Whitman Archive*.”
Have Digital Collections Led to Innovative Digital Scholarship?

- A few examples of digital scholarship associated with these digital collections:
  - Text mining: erotics of Dickinson (Plaisant, et al)
  - Interpretive exhibits: Uncle Tom’s Cabin conference
- But the digital collections themselves are probably the best examples of digital scholarship
II. What does it mean—and take—to produce digital scholarship?

http://www.pageflakes.com/lspiro/
The Dissertation Remix Project

- Pragmatic approach: The best way to explore digital scholarship is to produce it myself
- Remixing 2002 dissertation on bachelorhood in American literature and culture as a work of “digital scholarship”
- Objectives:
  - Determine feasibility of relying on digital resources for research
  - Experiment with tools for analyzing, visualizing, organizing, & mining information
  - Explore dissemination methods: blogs, wikis, multimedia essays
  - Reflect on challenges & opportunities for digital humanities
How feasible is it to rely on digital resources?

- Evaluating how many of my original resources are available electronically
- Assessing the quality of those sources
- Surveying what else is available online
- Using Zotero to capture, organize, analyze, & share research
### % of works on my bibliography digitized & available as full text (2008)

<table>
<thead>
<tr>
<th>Type</th>
<th>% Full Text</th>
<th>% in Digital Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>secondary monograph</td>
<td>23.5%</td>
<td>98.3%</td>
</tr>
<tr>
<td>secondary periodical</td>
<td>93.1%</td>
<td>93.1%</td>
</tr>
<tr>
<td>primary monograph</td>
<td>75.8%</td>
<td>97.0%</td>
</tr>
<tr>
<td>primary periodical</td>
<td>88.6%</td>
<td>91.1%</td>
</tr>
<tr>
<td>archival</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td><strong>Total Primary</strong></td>
<td>82.8%</td>
<td>91.9%</td>
</tr>
<tr>
<td><strong>Total Secondary</strong></td>
<td>37.2%</td>
<td>97.3%</td>
</tr>
<tr>
<td><strong>Grand Total</strong></td>
<td>59.1%</td>
<td>94.6%</td>
</tr>
</tbody>
</table>

[http://digitalscholarship.wordpress.com/2008/05/05/how-many-texts-have-been-digitized/](http://digitalscholarship.wordpress.com/2008/05/05/how-many-texts-have-been-digitized/)
What is the quality of digitized works? (subjective evaluation of 19th C books)

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Google Books</th>
<th>Open Content All.</th>
<th>EAF</th>
<th>Making of America</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scanning</td>
<td>☺☺☺</td>
<td>☺☺☺☺</td>
<td>☺☺☺</td>
<td>☺☺☺</td>
</tr>
<tr>
<td>Text conversion</td>
<td>☺☺</td>
<td>☺☺☺</td>
<td>☺☺☺</td>
<td>☺☺☺☺</td>
</tr>
<tr>
<td>Metadata</td>
<td>☺☺</td>
<td>☺☺☺</td>
<td>☺☺☺</td>
<td>☺☺☺☺☺</td>
</tr>
<tr>
<td>Terms of use</td>
<td>☺☺☺☺</td>
<td>☺☺☺☺--------------</td>
<td>☺</td>
<td>☺☺☺</td>
</tr>
<tr>
<td>Convenience</td>
<td>☺☺☺☺</td>
<td>☺☺☺</td>
<td>☺☺</td>
<td>☺☺☺</td>
</tr>
<tr>
<td>Reputation</td>
<td>☺☺</td>
<td>☺☺☺</td>
<td>???</td>
<td>☺☺☺☺☺</td>
</tr>
</tbody>
</table>

Surveying Google Books (GB) to see what I missed in my prior research

- Case study: the history of *Reveries of a Bachelor* (1850)
- Ran simple search in GB for “Reveries of a Bachelor”
- One result screen says: “101 - 150 of 690,” but then the very next one says “Books 151-159 of 159,” limiting what you can get access to

http://scholarship.rice.edu/handle/1911/21839
What I found in Google Books

- Publishing history
  - ads
  - publishers’ announcements
  - pricing
- History of reading
  - passages in memoirs
  - library catalogs
  - recitation scripts
- Intertextuality
  - books quoting or referencing *Reveries*

http://digitalscholarship.wordpress.com/2008/12/19/using-google-books-to-research-publishing-history/
Analyzing Information

Reveries as a word cloud
Using Text Visualization & Analysis to Expose Patterns & Feed Interpretation

- To what extent can I use software to stimulate & support interpretation?
- Experimenting with
  - Text visualization tools (e.g. Wordle)
  - Text analysis tools (e.g. TAPOR)
  - Text mining tools (e.g. MONK)*

*The next phase of my project
Using Software to Compare Reveries of a Bachelor to Pierre

- Can we use text analysis tools to study the relationship between texts?
- My notion: Melville’s Pierre is a bitter satire of Reveries of a Bachelor & other sentimental bachelor literature
- Used Wordle word cloud generator & TAPOR’s Comparator & collation tools to examine two works in relation to each each
Reveries Word Cloud (Wordle)
Pierre Word Cloud (Wordle)
Comparing Reveries & Pierre with Wordle
## Comparing *Reveries* & *Pierre* with TAPOR Comparator

<table>
<thead>
<tr>
<th>Words</th>
<th>Rev. counts</th>
<th>Rev relative</th>
<th>Pier relative</th>
<th>Pier counts</th>
<th>Rel ratio (R/P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mother</td>
<td>58</td>
<td>0.0009</td>
<td>0.0015</td>
<td>237</td>
<td>0.5953</td>
</tr>
<tr>
<td>father</td>
<td>39</td>
<td>0.0006</td>
<td>0.0009</td>
<td>138</td>
<td>0.6875</td>
</tr>
<tr>
<td>sweet</td>
<td>73</td>
<td>0.0011</td>
<td>0.0008</td>
<td>125</td>
<td>1.4206</td>
</tr>
<tr>
<td>light</td>
<td>45</td>
<td>0.0007</td>
<td>0.0007</td>
<td>106</td>
<td>1.0327</td>
</tr>
<tr>
<td>morning</td>
<td>56</td>
<td>0.0009</td>
<td>0.0005</td>
<td>86</td>
<td>1.584</td>
</tr>
<tr>
<td>night</td>
<td>68</td>
<td>0.001</td>
<td>0.0007</td>
<td>110</td>
<td>1.5037</td>
</tr>
<tr>
<td>dark</td>
<td>38</td>
<td>0.0006</td>
<td>0.0004</td>
<td>71</td>
<td>1.3019</td>
</tr>
<tr>
<td>time</td>
<td>106</td>
<td>0.0016</td>
<td>0.0014</td>
<td>227</td>
<td>1.1359</td>
</tr>
<tr>
<td>heart</td>
<td>199</td>
<td>0.003</td>
<td>0.0012</td>
<td>186</td>
<td>2.6026</td>
</tr>
<tr>
<td>hand</td>
<td>102</td>
<td>0.0016</td>
<td>0.001</td>
<td>163</td>
<td>1.5222</td>
</tr>
<tr>
<td>face</td>
<td>62</td>
<td>0.0009</td>
<td>0.001</td>
<td>162</td>
<td>0.931</td>
</tr>
<tr>
<td>eye</td>
<td>71</td>
<td>0.0011</td>
<td>0.0004</td>
<td>67</td>
<td>2.5778</td>
</tr>
<tr>
<td>love</td>
<td>134</td>
<td>0.002</td>
<td>0.0012</td>
<td>192</td>
<td>1.6977</td>
</tr>
<tr>
<td>think</td>
<td>70</td>
<td>0.0011</td>
<td>0.0005</td>
<td>85</td>
<td>2.0033</td>
</tr>
</tbody>
</table>
Words associated with “mother”:

<table>
<thead>
<tr>
<th>Reveries</th>
<th>Pierre</th>
</tr>
</thead>
<tbody>
<tr>
<td>heart</td>
<td>dear</td>
</tr>
<tr>
<td>kiss</td>
<td>conceal</td>
</tr>
<tr>
<td>lap</td>
<td>torture</td>
</tr>
</tbody>
</table>
Impact of Experiment with Text Analysis

- Allows you to extract out key features of texts
- But then you can recontextualize those features by using co-location, concordance & co-occurrence tools
- Establish a “linguistic profile”: see how Melville appropriates & twists language of sentimentality
- Reveals the dark undercurrents in Mitchell’s language
- Need to explore methodological issues:
  - Are the common words unique to these works?
  - How do you interpret word counts?
  - How would I use this information in an argument?
- Text analysis tools open up questions rather than provide definitive answers--stimulus to interpretation (cf. McGann, Ramsay, et al)
Disseminating Research

Image: http://www.flickr.com/photos/nic221/391536867/
The Traditional Model for Scholarly Communication

- Collect Information
- Analyze Information
- Shape Argument
- Disseminate Research
Speeding Up & Opening Up Scholarly Communication

- Collect Information
- Analyze Information
- Shape Argument
- Disseminate Research
- Share bibliography
- Share data
- Blog research
- Deposit in OA repository
Sharing Bookmarks & Research Collections*

http://www.diigo.com/user/lspiro/digital_humanities

* Soon I will use Zotero to share my research collections online.
Digital Scholarship in the Humanities Blog

Defining digital scholarship in the humanities:
Ten-fingered humanists?

Posted on November 30, 2007 by Lisa Spiro | Edit

“Digital scholarship” seems to have become a new buzzword in academia. The term is invoked by those advocating for open access to scholarly knowledge (e.g. Charles Bailey’s Digital Scholarship) as well as those promoting innovative research methodologies. Universities, libraries, and funding organizations are beginning to recognize the need to support digital scholarship. Witness:

- The opening of centers devoted to digital scholarship, such as the University of Richmond’s newly launched digital scholarship lab (supported by president Ed Ayers) and Johns Hopkins’s new Digital Research and Curation Center.
- New positions devoted to digital scholarship, such as NYU’s Librarian for Literary Studies and Digital Scholarship in the Humanities.
- Funding programs to support digital scholarship in the humanities, such as the NEH’s Digital Humanities Initiative and the ACLS Digital Innovation.

http://digitalscholarship.wordpress.com/
Why Blogging Has Been a Boon

- It enlarges my perspective
  - Comments from biologists & anthropologists as well as literary scholars & historians
  - Comments on my work from folks in UK, Spain, etc.
- My ideas have been challenged and improved through dialogue.
- I feel more engaged in the research community and more motivated.
- I have a well of ideas from which I can draw
- It increases the visibility of my work, and opens up thus more opportunities (to contribute essays, give presentations, review grants, etc.)

Image: http://www.flickr.com/photos/wakingtiger/3156791845/
Multimodal Scholarship: Marketing Marvel, the Movie

- Article on publishing history of *Reveries* limited because couldn’t present visual evidence: bindings, illustrations, etc.
- Why not turn article into short film that shows bibliographic features of different editions, as well as ads for them?
- Challenges:
  - Condensing core argument to 5 minute narrative
  - Thinking visually & cinematically
  - Citation practices for video?
  - Where to “publish”?
III. Technical Challenges (& Strategies) for Digital Scholarship

- Finding & using the appropriate tools
- Developing the necessary skills

There are, of course, many other challenges, such as tenure & promotion, funding, copyright, need for tools, etc. See Our Cultural Commonwealth.

Image: http://www.flickr.com/photos/jonlucas/204213403/
Finding the Right Tools

- There are hundreds of tools relevant for research--e.g. tools for creating bibliographies, performing text analysis, writing collaboratively, etc.
- How do you find such tools and figure out which are best?
- Why not have a web site that categorizes & reviews research tools?
Types of Tools

I want to...

- Analyze statistics
- Analyze texts
- Author an interactive work
- Blog
- Brainstorm/ generate ideas
- Build and share collections
- Collaborate
- Collect data
- Communicate with colleagues
- Compare resources
- Conduct linguistic research
- Convert/ manipulate files

### Design Principles for DiRT

- Make it a wiki, so anyone can contribute
- Organize it clearly, based on what researchers want to do
- Furnish clear criteria for evaluation
- Be flexible. Evolve wiki according to community needs.
Why I Need to Learn to Program

- My confession: I’m a digital humanist with only limited programming skills (Perl & XSLT)
- Enhancing my programming skills would allow me to:
  - Avoid so much tedious, manual work
  - Do citation analysis
  - Pre-process texts (remove the junk)
  - Automatically download web pages
  - And much more…
OMG, I Need to Learn Math

- I don’t know this language:
  \[ H(P) = -\sum_{i=1}^{n} p(s_i) \times \log(p(s_i)) \]
  Claude Shannon’s Entropy Formula (I think)

- Doug Oard: “humanities scholars are going to need to learn a bit of probability theory”

- Sculley & Pasanek, “Meaning and mining”
  - 4 different experiments using 4 algorithms to test classification of metaphors in 18th C political discourse yield 4 different results.
  - Suggests best practices for humanities data mining, e.g. being explicit, making data available, peer review of method
How Humanists Can Confront the Skill Challenge

- Collaborate with computer scientists, linguists, librarians, etc. to develop tools & scholarship

- Develop new skills:
  - Digital Humanities Institutes (e.g. NEH, U of Victoria)
  - Self-instruction (could be done in study groups)
    - Bill Turkel & Alan McEarchen, *The Programming Historian*
  - Skill-building sabbaticals & fellowships
  - Incorporate training into graduate programs
  - Just do it!
Why Digital Scholarship in the Humanities Is Worth the Effort

- It’s fun—I’m always learning
- It’s typically collaborative
- It seems relevant
- It reaches an audience beyond academia
- You can accomplish what would be difficult to do without digital technologies
- It has the potential to advance humanities research
- Humanities scholars need to engage with the information age (cf. Davidson)
Bonus Slides
Does Digital Humanities Foster a Renewed Focus on Method?

Are we entering a “new phase of scholarship that will be dominated not by ideas, but once again by organizing activities, both in terms of organizing knowledge and organizing ourselves and our work”?

“The Scientific Method”

Research Methods in the Social Sciences

1. Formulating a research problem
2. Conceptualizing a research design
3. Constructing an instrument for data collection
4. Selecting a sample
5. Writing a research proposal
6. Collecting data
7. Processing data
8. Writing a research report

Ranjit Kumar, Research Methodology
Typical Literary Research Method

1. Identify research question
2. Define theoretical approach(es)
3. Find & gather relevant primary & secondary sources
4. Read & annotate sources
5. Develop an interpretation & write the paper/book
6. Publish the work
## Literary Research in a Digital Environment

<table>
<thead>
<tr>
<th>Research Process</th>
<th>Possible Implications of the Digital</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify research question</td>
<td>Collaborative research, global humanistic studies, media studies</td>
</tr>
<tr>
<td>Gather sources</td>
<td>Search engines, mass digitization, RSS feeds, recommendation services, etc</td>
</tr>
<tr>
<td>Read &amp; annotate</td>
<td>“Distant reading,” “not reading,” annotation tools (Zotero, Pliny)</td>
</tr>
<tr>
<td>Analyze &amp; interpret</td>
<td>Text mining, analysis, visualization</td>
</tr>
<tr>
<td>Disseminate research</td>
<td>Blogs, multimodal scholarship, peer-to-peer review, open access repositories</td>
</tr>
</tbody>
</table>
Why should humanities engage with computing?

- Jerome McGann: “Because we have no choice.” The archive is becoming digital.
- Cathy Davidson: “Hybridity, exchange, flow, and cultural transaction are all explored more responsibly and adventurously when the resources of many nations, in many languages, have been digitized, made interoperable, and offered for research by scholars around the world.”
- Brett Bobley of the NEH: “it is about getting things done that couldn’t be done before.”