DHSI

DIGITAL HUMANITIES SUMMER INSTITUTE

Conceptualising and Creating a Digital Edition

Jennifer Stertzer, Cathy Hajo, and Erica Cavanaugh

ETCL
ELECTRONIC TEXTUAL CULTURES LAB

Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada

Welcome to DHSI 2023!

Thank you for joining the DHSI community!

In this coursepack, you will find essential workshop materials prefaced by some useful general information about DHSI 2023.

Given our community's focus on things computational, it will be a surprise to no one that we might expect additional information and materials online for some of the workshops—which will be made available to you where applicable—or that the most current version of all DHSI-related information may be found on our website at dhsi.org. Do check in there first if you need any information that's not in this coursepack.

Please also note that materials in DHSI's online workshop folders could be updated at any point. We recommend checking back on any DHSI online workshop folder(s) that have been shared with you in case additional materials are added as DHSI approaches and takes place.

And please don't hesitate to be in touch with us at **institut@uvic.ca** or via Twitter at **@AlyssaA_DHSI** or **@DHInstitute** if we can be of any help.

We hope you enjoy your time with us!

DHSI
DIGITAL HUMANITIES SUMMER INSTITUTE

ETCL
ELECTRONIC TEXTUAL CULTURES LAB

Statement of Ethics & Inclusion

Please review the DHSI Statement of Ethics & Inclusion available here: **https://dhsi.org/statement-of-ethics-inclusion/**

DHSI is dedicated to offering a safe, respectful, friendly, and collegial environment for the benefit of everyone who attends and for the advancement of the interests that bring us together. There is no place at DHSI for harassment or intimidation of any kind.

By registering for DHSI, you have agreed to comply with these commitments.

Virtual Sessions

Your registration in DHSI 2023 also includes access to the virtual **institute lecture** sessions. Access details for these talks will be shared as DHSI approaches.

Due to the high volume of attendees, please ensure your DHSI registration name or DHSI preferred name and your Zoom name match so that we know to let you into the virtual sessions.

DHSI Materials

DHSI materials (ex. videos, documents, etc.) are intended for registrant use only. By registering, you have agreed that you will not circulate any DHSI content. If someone asks you for the materials, please invite them to complete the registration form to request access or contact us at **institut@uvic.ca**.

Auditor and participant registration

If you registered to **audit** any workshops, note that auditor involvement is intended to be fully self-directed without active participation in the workshop. The auditor option offers more flexibility regarding pace and time with the workshop content. Your registration as an auditor will include access to some asynchronous workshop materials only and does not include access to live workshop sessions and/or individual/group instruction or consultation. Please direct any questions about DHSI workshop auditing to **institut@uvic.ca**.

If you registered as a **participant** in any workshops, your registration includes access to asynchronous content + active participation in live workshop session(s). The workshop instructor(s) will contact you about the date(s), time(s), and platform(s) of the live workshop session(s).

If you are unsure whether you registered as an auditor or participant, please check your registration confirmation email. Further questions can be directed to **institut@uvic.ca**.

Schedule

The at-a-glance schedule of DHSI 2023 courses, workshops, institute lectures and aligned conferences & events can be found here:
https://dhsi.org/timetable/

All times are listed in North American **Pacific Time Zone**.

For those who registered as participants in any workshops, live sessions for online workshops are not currently listed on the above-referenced schedule. **Instructors will be in touch with registered participants directly about the exact date(s) and time(s) of their live workshop session(s).**

Acknowledgements

We would like to thank our partners and sponsors (including the Social Sciences and Humanities Research Council), workshop instructors, aligned conference & event organizers, institute lecturers, local facilitators, and beyond for making this possible.

Further information

General DHSI 2023 information: **https://dhsi.org/program/**

Full course listings (in-person): **https://dhsi.org/on-campus-courses/**

Full workshop listings (online): **https://dhsi.org/online-workshops/**

Aligned conferences & events (in-person): **https://dhsi.org/on-campus-aligned-conferences-events/**

Aligned conferences & events (online): **https://dhsi.org/online-aligned-conferences-events/**

Institute lectures: **https://dhsi.org/institute-lectures/**

Frequently asked questions: **https://dhsi.org/faq/**

Any questions not addressed in the above pages? Please email us at **institut@uvic.ca**!

Conceptualising and Creating a Digital Edition - Coursepack

Instructors: Cathy Moran Hajo, Erica Cavanaugh, and Jennifer Stertzer

This course will explore all aspects of conceptualizing, planning for, and creating a digital edition. It provides a basic introduction to the various types of digital editions, the practice of editing in the digital age, and a survey of the many digital tools available to serve project goals. Approaching a digital edition means taking time to think about how end-users will want to work with a particular edition. Beginning with the research and analytical needs of end-users in mind, editors are better able to develop effective editorial strategies that will result in dynamic, accessible, and functional digital editions. In this course, participants will engage in hands-on learning and group discussions related to project conceptualization, editorial policies and processes, and the selection and use of digital tools that can serve the needs of researchers and other end-users. Participants will bring a selection of sample materials they are working with so that they can experiment with methods and tools during the week. Our goal is for participants to return to their home institutions ready and able to build upon, enhance, and transform these initial ideas into robust digital editions.

What to Bring

Bring copies of your documents/items to work on, about 10-20 pages.

- Select representatives of the more difficult and complex documents that you have
- Select a broad range of the kinds of text formats that you want to include
- If you have transcriptions of the texts, bring them.

A laptop computer

Schedule

***Please note, each session will begin with an overview of the topics. We will then, as a class, decide what topics to explore and discuss in-depth. We will also tailor the hands-on time to meet the needs of the participants.**

Day 1

Morning – What is an electronic/online/digital edition? Features of a digital edition and how they differ from print editions. Planning and conceptualizing a digital edition: we will consider project mission and goals, and examine several digital edition categories and methodical frameworks.

Late morning / Early Afternoon – Participants share the current state of their projects, their end-user goals, and working plans for their digital edition, as well as discuss the kinds of documents/items they are working with. The goal of this session is to help participants consider

key aspects of project management and workflow strategies.

Afternoon – Overview of digital editions. We will examine the features, benefits, and drawbacks of each type. We'll also explore the intersection of digital humanities and digital editing.

Day 2

Morning/Afternoon – Editorial processes and policies (searching, imaging, transcribing, editing, annotation, indexing, etc.) will be examined, with consideration of how editorial decisions are both informed by and influence digital publication and end-user goals.

Day 3

Morning/Afternoon – Overview of tools and technologies. Examination of available content management systems/platforms, tools, and workflows to handle every aspect of the editorial and conversion/digitization process (i.e. acquisition, permissions, transcription, proofing, annotation, fact checking, copy editing, manuscript review, digital publication). Discussion of available tools/platforms: Word, XML, oXygen, FairCopy, Omeka, Drupal, Scalar, and WordPress. We will also discuss born digital versus legacy conversion and how this will affect options. And finally, we will consider several key issues related to digital publication (interface development and use of consultants, hosting, etc.).

Resources:

- oXygen [http://www.oxygenxml.com]
- FairCopy [https://www.performantsoftware.com/projects/free-early-access-now-faircopy/]
- Omeka [https://omeka.org]
- Drupal [https://drupal.org]
- Scalar [http://scalar.usc.edu]

Day 4

Morning – How to evaluate the goals of the project, ideal workflow and editorial environment, publication/user interface plan and make decisions about tools and platform. Discussion of how to plan, design, organize, layout, and manage your edition and site.

Afternoon – From paper to screen - an opportunity to use Omeka, Drupal, or other digital tools for your project.

Day 5

Course wrap-up as well as additional hands-on time. Discussion of what the future holds for digital documentary editions as well as the newest developments in visualizations, sound,

images, and moving images.

Suggested Readings:

- *Digital Scholarly Editing: Theories and Practices*, edited by Matthew James Driscoll and Elena Pierazzo (Cambridge, 2016)
    - This volume presents the state of the art in digital scholarly editing. Drawing together the work of established and emerging researchers, it gives pause at a crucial moment in the history of technology in order to offer a sustained reflection on the practices involved in producing, editing and reading digital scholarly editions-and the theories that underpin them.
- *Publishing Scholarly Editions: Archives, Computing, and Experience*, by Christopher Ohge (Cambridge, 2021)
    - Publishing Scholarly Editions offers new intellectual tools for publishing digital editions that bring readers closer to the experimental practices of literature, editing, and reading. After the Introduction (Section 1), Sections 2 and 3 frame intentionality and data analysis as intersubjective, interrelated, and illustrative of experience-as-experimentation. These ideas are demonstrated in two editorial exhibitions of nineteenth-century works: Herman Melville's Billy Budd, Sailor, and the anti-slavery anthology The Bow in the Cloud, edited by Mary Anne Rawson. Section 4 uses pragmatism to rethink editorial principles and data modelling, arguing for a broader conception of the edition rooted in data collections and multimedia experience. The Conclusion (Section 5) draws attention to the challenges of publishing digital editions, and why digital editions have failed to be supported by the publishing industry. If publications are conceived as pragmatic inventions based on reliable, open-access data collections, then editing can embrace the critical, aesthetic, and experimental affordances of editions of experience.
- *A Guide to Documentary Editing*, Third Edition, by Mary-Jo Kline and Susan Holbrook Perdue (Virginia, 2008)
    - This guide provides a deep understanding of the history and practice of scholarly editing, familiarity with the terminology editors employ, and advice on the major decisions to be made. It is chiefly addressed towards the creation of traditional print editions.
- *Editing Historical Documents: A Handbook of Practice*, Michael E. Stevens and Steven B. Berg (AltaMira Press, 1997)
    - This book provides examples of practice from a variety of different editing projects and covers all the major editing tasks--transcription, annotation, indexing, etc.

- Cameron Blevins, "Martha Ballard's Diaries," (4 blog posts), Cameron Blevins, 2009-2010.
- Quinn Dombrowski, "Choosing a platform for your project website" [http://digitalhumanities.berkeley.edu/blog/13/12/04/choosing-platform-your-project-website]
- Andrew Jewell, "Digital Editions: Scholarly Tradition in an Avant-Garde Medium," Annual Meeting, Association for Documentary Editing, 2008. [http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1193&context=libraryscience]
- MLA, Guidelines for Editors of Scholarly Editions. [https://www.mla.org/Resources/Guidelines-and-Data/Reports-and-Professional-Guidelines/Publishing-and-Scholarship/Guidelines-for-Editors-of-Scholarly-Editions]
- MLA Statement on the Scholarly Edition in the Digital Age. [https://www.mla.org/content/download/52050/file/rptCSE16.pdf]
- Kenneth M. Price, "Edition, Project, Database, Archive, Thematic Research Collection: What's In a Name?," DHQ: Digital Humanities Quarterly. Summer 2009 3:3. [http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1068&context=englishfacpubs]

Other Resources:

- Scholarly Editing: The Annual of the Association for Documentary Editing
    - *Scholarly Editing* is an open-access, peer-reviewed journal committed to the development and advancement of all aspects of textual and documentary editing, including the recovery of texts and artifacts that represent and celebrate the lives and contributions from and about Black, Latinx, and Indigenous peoples; Asian Americans and Pacific Islanders; women; LGBTQ+ individuals; and peoples and cultures of the Global South. In addition to projects that illustrate the traditional range of editorial methodologies and practices, we welcome those that feature rare or marginal texts, texts that dislodge the single-author model, oral histories and tales, community recovery, creative works of "rememory," and the decolonizing of artistic works, archives, records, and editions for the discoverability of racialized and underrepresented stories and cultural artifacts.
- eLaboratories
    - eLaboratories—or eLabs—is an emergent space for cultivating connections, conversations, and collaborations between a diverse community of practitioners engaged in editing, recovery, or other research activities related to making source materials accessible and discoverable. Through our courses, forums, events, and more, practitioners can share their passions, grow their expertise, and shape the discourse of a vibrant, innovative, and inclusive community.

August 31, 2009 by Cameron Blevins

Text Analysis of Martha Ballard's Diary (Part 1)

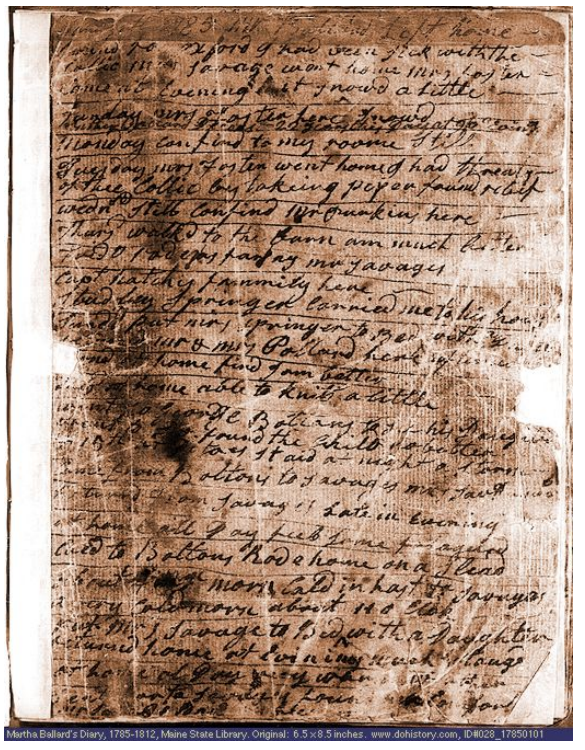
"mr Ballard left home bound for Oxford. I had been Sick with the Collic. mrs Savage went home. mrs foster Came at Evening. it snowd a little."

This is the first entry in the diary of Martha Ballard. Martha Ballard was a rural Maine midwife who kept an extensive diary between 1785 and 1812 and whose life was immortalized in 1990 by the historian Laurel Thatcher Ulrich's award-winning *A Midwife's Tale*. Over the course of three decades, Ballard kept a meticulous, near-daily accounting of her life spanning over 10,000 entries.

When reading *A Midwife's Tale*, I was struck by how readily the text would seem to lend itself to digital analysis. In an interview, Ulrich noted, "The very thing that had attracted me to the diary in the first place was also the thing that made it difficult to work with. I mean there's just so much." To ground herself, she began by simply counting things: "And I would go day by day for every other year of the diary, and I would tick off what was in each entry: baking or brewing, spinning or washing, or trading, sewing, mending, deliveries, general medical accounts, going to church, visitors, people coming for meals, etc." Because of the sprawling scope, she took this quantitative approach only for the even-numbered years in the diary. The fact that she was working in the late eighties without a computer

makes her work even more impressive.

After poking around online I came across DoHistory.org, a website developed and maintained by the Film Study Center at Harvard University and hosted by (who else, really) George Mason's CHNM. The website presents the diary to the public in two formats: the viewer can either browse through photographed pages of the diary or read the transcript of the pages (transcribed through a monumental effort by Robert R. McCausland and Cynthia MacAlman McCausland):



[1]	[7]	mr Ballard left home bound for Oxford. I had been Sick with the Collic. mrs Savage went home. mrs foster Came at Evening. it snowd a little.	
[2]	[1]	Sunday. mrs Foster here. Snowd. [Father] Ballard Deceast 22 years this Day, at 9/o^{c} Evin^{g}.	
[3]	[2]	Monday. Confined to my roome Still.	
[4]	[3]	Tuesday. mrs. Foster went home. I had threats of thee Collic; by takeing peper found releif.	
[5]	[4]	Wedn^{sd} Still Confind. mr Purkins here.	
[6]	[5]	Thurs^{d}. walk^{d} to the barn. am much better.	
[7]	[6]	Frid. I rode as far as mr Savages.	
[8]	[7]	Capt hatchis fammily here. Studley Springer Carried me to his hous.	
[9]	[1]	Sund. I put mrs springer to Bed with [a daughter]. [returnd]. mr & mrs Pollard here afternoon.	
[10]	[2]	Mond. at home. find I am better.	
[11]	[3]	Tuesd. at home. able to knit a little.	
[12]	[4]	Went to George Boltans to see his Daughter.	
[13]	[5]	Went to Ditos. found the Child no better.	
[14]	[6]	14 & 15th at Ditoes. staid. at night a storm.	
[16]	[1]	Came from Boltons to Savages. mrs Sav^{g} unwell.	
[17]	[2]	Returnd From Savages Late in Evening.	
[18]	[3]	At home all Day. feel Some fatagued.	
[19]	[4]	Cald to Boltons. Rode home on a Slead.	
[20]	[5]	About [ ] morn. Cald in hast to Savages. a very Cold morn. about 11 o clok put mrs Savage to Bed with a Daughter. Returnd home at Evn very much fatagued.	
[21]	[6]	at home all Day. very warm weather.	
[22]	[7]	Went [ ] Savages. found her Comfortable. mr Savage Paid [ ].	

When I realized the entire diary was online, it got me thinking about possibilities for text mining. As an aspiring digital humanist with little "hard" skills beyond basic GIS, I had been meaning to learn how to program for quite some time. In Martha Ballard's diary, I had an intriguing source of data with which to learn how to do so. Now I just had to learn how to program. With the patient help of several programming-savvy family members, I gradually learned the basics of Python and how to apply it to Martha Ballard's diary. What follows are the first steps we took to process the diary's raw data into an accessible digital format.

Process

At first, I briefly considered learning how to scrape the text of the diary off the

website. After some investigation, I decided that was a little beyond my abilities, so I copped out to the much easier route of sending an email to Kelly Schrum at CHNM, who kindly forwarded my request to Ammon Shepherd, who emailed me a zip file containing 1,431 html documents, one for each page of the diary. The html files of the transcribed diary are a basic, 3-column table that look this. My first step was to find a way to strip out the html tags and organize the text into a systematic database of individual entries. Fortunately, Ballard's meticulousness and consistency lent itself well to such an approach.

The diary's format translates quite nicely into creating a list of lists – the "main" diary being a list of all the entries, and each entry being a list in and of itself. The first program we wrote was to open each html file and begin extracting the different sections of text (which were conveniently marked by html tags). Iterating through each entry allowed us to separate the different columns in her diary into different items in the list. Here is the breakdown of our "list of lists":

1. **Diary**
    1. **Entry**
        1. **Date**
            1. **Month**
            2. **Day**
            3. **Year**
        2. **Day of the Week**
        3. **Main Text of Entry**
        4. **Day Summaries** (Column 3 of actual diary entry)
        5. **Birth(s)** (Recorded in Column 1 of actual diary entry)

In creating the list, we had to separate out the raw data from the html tags that

formatted it. Fortunately, the folks who built the html files originally used an extremely systematic formatting process that actually made the job of distilling one from the other quite straightforward. A Python module called Pickle allowed us to export the list of entries as a manageable single file that we could then easily import into future programs to manipulate.

For example, the third entry in the diary would translate a bit into something like this:

1. **Diary**
    1. **Entry (3)**

        1. **Date**
            1. **1** (January)
            2. **3**
            3. **1785**
        2. **3** (Tuesday – Ballard numbered the weekdays, beginning with Sunday as 1)
        3. **“Tuesday. mrs. Foster went home. I had threats of thee Collic; by takein peper found releif.”**
        4. Empty
        5. Empty

The list allows us to access pieces of information by “calling” their position. It helped me to think of the entire diary list as a warehouse containing almost 10,000 boxes (entries) inside it, with each box containing five compartments, with the first of *those* compartments divided into three sub-compartments. If you were to open any of the boxes (entries) and look inside the first compartment, then inside sub-compartment number two, you would always find a number that

represented the month of that particular entry. If you were to look inside the third compartment of the entry/box, you would always find the main text for that day's entry.

The advantages of setting up the data in a list structure is the ability to access these specific pieces of information easily and to compare them across entries. In many ways, processing the text to make it readable and programmable is one of the biggest challenges to text mining. Deciding on the most logical way to organize and break down over 1,400 files will lay the groundwork for the fun part: writing programs to actually analyze the diary of Martha Ballard.

****Special-edition sneak preview of future posts in this series****

A simple counting program reveals that the main text of Martha Ballard's diary *alone* contains **377,315** words, spanning I-couldn't-make-this-number-up **9,999** entries. That is a lot of data to play with.

#A Midwife's Tale #Laurel Ulrich #Martha Ballard #Programming #Python #Text analysis

---

Comments

Ben Brumfield - August 31, 2009 @ 9:48 am

I'm looking forward to see what you come up with. I've made a couple of stabs at mining diary data in a naive way on my own project, but suspect that my approach is too strongly influenced by the text I'm working with.

- **Cameron Blevins - August 31, 2009 @ 11:28 pm**

  Ben,

  Thanks for the comment – I think it's pretty normal for methodological approaches to get shaded by the text, which I admit can be both a good and bad thing. From what I understand you have a far

more sophisticated grasp of programming than I do, so I welcome any and all advice from your own experience.
-Cameron

Ben Brumfield - August 31, 2009 @ 11:59 pm

That's a bit comforting. I'd love to chat with you about this, but suggest that we wait until you've posted your observations. I'm very interested to see what sort of data you think is extractable.

Kelly in Kansas - September 3, 2009 @ 7:26 am

Thanks for letting us follow your interesting journey virtually. We will all learn from your experience.

- **McCausland - July 6, 2010 @ 10:46 am**
  You boys have taken on one h— of a job
  and I for one would consider it a personal favor if
  you would alert me when you have this project well under way.
  I'm the guy (with Cyn) that transcribed the 9,999 words that Martha wrote.
  Robert McCausland

  - **Cameron Blevins - July 7, 2010 @ 8:56 am**
    Thanks Rob! Good to hear from you again. I'll definitely be keeping you posted.
    -Cameron

Leave a Reply

Your email address will not be published / Required fields are marked *

Name*

Email*

September 9, 2009 by Cameron Blevins

Text Analysis of Martha Ballard's Diary (Part 2)

Given Martha Ballard's profession as a midwife, it is no surprise that she carefully recorded the 814 births she attended between 1785 and 1812. These events were given precedence over more mundane occurrences by noting them in a separate column from the main entry. Doing so allowed her to keep track not only of the births, but also record payments and restitution for her work. These hundreds of births constituted one of the bedrocks of Ballard's experience as a skilled and prolific midwife, and this is reflected in her diary.

As births were such a consistent and methodically recorded theme in Ballard's life, I decided to begin my programming with a basic examination of the deliveries she attended. This examination would take the form of counting the number of deliveries throughout the course of the diary and grouping them by various time-related characteristics, namely: year, month, and day of the week.

Process and Results

The first basic step for performing a more detailed text analysis of Martha Ballard's diary was to begin cleaning up the data. One step was to take all the words and (temporarily) turn every uppercase letter into a lowercase letter. This

kept Python from seeing “Birth” and “birth” as two separate words. For the purposes of this particular program, it was more important to distill words into a basic unit rather than maintain the complexity of capitalized characters.

Once the data was scrubbed, we could turn to writing a program that would count the number of deliveries recorded in the diary. The program we wrote does the following:

1. Checks to see if Ballard wrote anything in the “birth” column (the first column of the entries that she also used to keep track of deliveries)
2. If she did write anything in that column, check to see if it contains any of the words: “birth”, “brt”, or “born”.
3. I then printed the remainder of the entries that contained text in the “birth” column but did not contain one of the above words. From this short list I manually added an additional seven entries into the program, in which she appeared to have attended a delivery but did not record it using the above words.

Using these parameters, the program could iterate through the text and recognize the occurrence of a delivery. Now we could begin to organize these births.

First, we returned the birth counts for each year of the diary, which were then inserted into a table and charted in Excel:

At the risk of turning my analysis into a John Henry-esque woman vs. machine, I compared my figures to the chart that Laurel Ulrich created in *A Midwife’s Tale* that tallied the births Ballard attended (on page 232 of the soft-cover edition). The two charts follow the same broad pattern:

Note: I reverse-built her chart by creating a table from the printed chart, then making my own bar graph. Somewhere in the translation I seem to have misplaced one of the deliveries (Ulrich lists 814 total, whereas I keep counting 813 on her graph). Sorry!

However, a closer look reveals small discrepancies in the numbers for each individual year. I calculated each year's discrepancy as follows, using Ulrich's numbers as the "true" figures (she is the acting President of the AHA, after all) from which my own figures deviated, and found that the average deviation for a given year was 4.86%. Apologies for the poor formatting, I had trouble inserting tables into WordPress:

Year	Deliveries Count		Difference	Deviation (from Ulrich)
	Manual (Ulrich)	Computer Program		
1785	28	24	4	14.29%
1786	33	35	2	6.06%
1787	33	33	0	0.00%
1788	27	28	1	3.70%
1789	40	43	3	7.50%
1790	34	35	1	2.94%
1791	39	39	0	0.00%
1792	41	43	2	4.88%
1793	53	50	3	5.66%

1794	48	48	0	0.00%
1795	50	55	5	10.00%
1796	59	56	3	5.08%
1797	54	55	1	1.85%
1798	38	38	0	0.00%
1799	50	51	1	2.00%
1800	27	23	4	14.81%
1801	18	14	4	22.22%
1802	11	12	1	9.09%
1803	19	18	1	5.26%
1804	11	11	0	0.00%
1805	8	8	0	0.00%
1806	10	11	1	10.00%
1807	13	13	0	0.00%
1808	3	3	0	0.00%
1809	21	22	1	4.76%
1810	17	18	1	5.88%
1811	14	14	0	0.00%
1812	14	14	0	0.00%

Keeping the knowledge in the back of my mind that my birth analysis differed slightly from Ulrich's, I went on to compare my figures with other factors, including the frequency of deliveries by month over the course of the diary.

If we extend the results of this chart and assume a standard nine-month pregnancy, we can also determine roughly which months that Ballard's neighbors were most likely to be having sex. Unsurprisingly, the warmer period between May and August appears to be a particularly fertile time:

Finally, I looked at how often births occurred on different days of the week. There wasn't a strong pattern, beyond the fact that Sunday and Thursday seemed to be abnormally common days for deliveries. I'm not sure why that was the case, but would love to hear speculation from any readers.

Analysis

The discrepancies between the program's tally of deliveries and Ulrich's delivery count speak to broader issues in "digital" text mining versus "manual" text mining:

Data Quality

Ulrich's analysis is a result of countless hours spent eye-to-page with the original text. And as every history teacher drills into their students when conducting research, looking directly at the primary documents minimizes the degrees of interpretation that can alter the original documents. In comparison, my analysis is the result of the original text going through several levels of transformation, like a game of telephone:

Original text -> Typed transcription -> HTML tables -> Python list -> Text file -> Excel table/chart

Each level increases the chance of a mistake. For instance, a quick manual examination using the online version of the diary for 1785 finds an instance of a delivery (marked by 'Birth') showing up in the online HTML, but which does not appear in the "raw" HTML files our program is processing and analyzing.

On the other hand, a machine doesn't get tired and miscount a word tally or accidently skip an entry.

Context

Ulrich brings to bear on the her textual analysis years of historical training and experience along with a deeply intimate understanding of Ballard's diary. This allows her to take into account one of the most important aspects of reading a document: context. Meanwhile, our program's ability to understand context is limited quite specifically to the criteria we use to build it. If Ballard attended a delivery but did not mark it in the standard "birth" column like the others, she might mention it more subtly in the main body of the entry. Whereas Ulrich could recognize this and count it as a delivery, our program cannot (at least with the current criteria).

Where the "traditional" skills of a historian come into play with data mining is in the arena of *defining* these criteria. Using her understanding of the text on a traditional level, Ulrich could create far, far superior criteria than I could for counting the number of deliveries Martha Ballard attends. The trick comes in translating a historian's instinctual eye into a carefully spelled-out list of criteria for the program.

Revision

One area that is advantageous for digital text mining is that of revising the program. Hypothetically, if I realized at a later point that Ballard was also tallying births using another method (maybe a different abbreviated word), it's fairly simple to add this to the program's criteria, hit the "Run" button, and immediately see the updated figures for the number of deliveries. In contrast, it would be much, much more difficult to do so manually, especially if the realization came at, say, entry number 7,819. The prospect of re-skimming thousands of entries to update your totals would be fairly daunting.

#A Midwife's Tale #Digital History #Laurel Ulrich #Martha Ballard #Programming #Python #Text analysis

Comments

Ben Brumfield - September 13, 2009 @ 2:05 pm

What a great idea to perform the same analysis (birth count by year) that Ulrich did, and compare your results to hers! I'm curious if you tried the manual method for the years with the widest divergence to see exactly why your program disagreed with Ulrich 5 times in 1795. That might offer an opportunity to improve the algorithm, but would more likely illustrate the limitations of data mining via text searches, with some concrete examples of why some analysis is non-computable.

Given the content of the diary, I wonder if you could look for correlations between births and other events that Ballard mentioned in the text of her entries. For example, she mentions the weather in the early pages I've examined, so you might be able parse out her descriptions (looking for strings like "fine" or "snow"), assign weather values to dates, then look for correlations between the weather and the deliveries she attended. Other events may be harder to identify: Ballard mentions her own health, but also comments on other people who are unwell. This might make it impossible to correlate deliveries to Ballard's health.

Thanks for posting such a detailed description of your work.

- **Cameron Blevins - September 14, 2009 @ 9:02 am**

Ben,

Interesting ideas – comparing my program to Ulrich's analysis certainly reinforced some of the limitations of data mining, but gave me hope in that it's not so difficult to tweak the program and make it more effective.

I really like the idea of looking for correlations between births and other events. I think the next step for an in-depth and systematic analysis of the text would be to create first a dictionary of unique words, then start grouping the words together under different categories (Religion, Death, Marriage, etc.). From there it would be really cool to then look for patterns using those groupings. Unfortunately Ballard's unique spelling system presents a challenge – she spells each word about 3-4 different ways, and has an incredible use of shorthand that contributes to around 37-38,000 "unique" words that would need to be cataloged. But if that gets done, the possibilities really become endless.

Thanks again for the support!

-Cameron

- **Ben Brumfield - September 14, 2009 @ 8:53 pm**
  I've encountered similar challenges editing the Julia Brumfield diaries, where proper names are spelled inconsistently — sometimes within the same page. Because I'm identifying terms for indexing/analysis as I transcribe the text from scanned images, I can resolve the spelling irregularities during transcription/editing. However, I've still found that full-text searches will identify terms I missed during the mark-up phase, so I can't say that my technique for data extraction is substantially better.

  I like your idea of extracting words from the text to identify variant spellings. I presume you'd do a frequency count over the entire corpus and sort the word/count pairs alphabetically to look for variants.

  Another possibility is to manually pull all the variant spellings of something you're interested in (say, weather) from one year's worth of entries. You could then execute the search against a different year and then manually identify missed variants there to see how representative your orgininal sample was. Someone with more statistics than I command could probably come up with reasonable figures for your extraction algorithm's accuracy. At any rate, you'd then be able to extract the data about that subject for your analysis.

  The downside of my approach is that you can block yourself from discovering subjects to investigate. Because you set out with one topic of analysis in mind (weather, say), you might miss the sort of things that a high-frequency word list could suggest. In my own project, I did not identify clothes washing as a domestic activity worthy of analysis until I was around 500

pages in. Fixing this will not be easy.

Larry H Cebula - September 14, 2009 @ 5:57 pm

Thank you for an interesting and helpful post. And yet is seems to confirm my fairly uninformed and perhaps knee-jerk reaction to a lot of these text mining projects–the conclusions are quite modest compared to the effort that went into the process. You have shown that Mainers had sex more often in the summer. No, you have shown that Mainers who hired Martha Ballard to midwife their babies were more likely to have had sex in the summer.

I hope I don't come off as snarky, I truly am impressed by your technical abilities. I think that is exactly what makes me read your conclusions and say "Is that it?"

- **Ben Brumfield - September 14, 2009 @ 7:17 pm**

  Larry, I'm afraid that you're confusing the technique (parsing and extracting data from the text) with the analysis (what you do with the data you've extracted, and whether you attempt to extract more data from the text). In this case, Cameron's extracted births and dates from the text — a single fact (Ballard's attendance at births) with a single dimension (the date the birth occurred). There's not very much analysis he can perform on this data, since there are only a limited set of questions to be asked from it. So of course the conclusions are modest.

  However, I'd wager that those conclusions–the graph of births per year–were determined through far less effort by Cameron than the effort spent by Ulrich to manually tabulate births by year. The fact that he's able to compare his low-cost effort to Ulrich's and such minor deviation lets us know the quality-to-cost trade-offs of his methodology. That in itself is worth knowing.

  The real question is–having exhausted the interesting questions he can ask of the data he's extracted–is there other data to be extracted from this text that might lend itself to more interesting analysis? How often was Ballard paid, and how many clients stiffed her? What were the geographic limits of her practice? If you had the misfortune to enter labor during a snowstorm, did that reduce the likelihood that you'd be attended by a midwife? If so, does weather explain the trough in births Ballard records between November and January, thus canceling out the summer-conception effect Cameron's initial analysis finds?

  Some of these may simply not be extractable from the kind of full-text search Cameron's performing here — geography in particular requires contextual information about where her clients lived that is not internal to the text, and which we're unlikely to have elsewhere. But it's a useful exercise to figure out which of these questions are answerable, which are impossible, and why.

- **Cameron Blevins - September 14, 2009 @ 10:45 pm**

  Larry,

Thanks for the feedback, you raise some good points. One fundamental issue with text mining in the humanities has been a gulf between promise and delivery – there seems to be so many things that could potentially be done, but that in the end prove to be either impossible or involve even more work than doing it by hand. There's also the issue of what I believe you termed "parlour tricks" on your blog, of analysis that may be superficially interesting or catchy, but adds little substantive value to the investigation. Both of these are fair criticisms.

In response to "Is that it?", I'd say its a valid question to ask since the analysis I've done so far isn't particularly deep, but that it's a bit like watching a ten year old learning how to play basketball and saying "Okay, but can they dunk?" Much like a ten year old struggling to learn how to play a new sport, the process for me (admittedly somewhat selfishly) has been more about the learning experience than about producing earth-shattering results.

Having said that, even my limited experience so far has affirmed for me the potential and ability of text mining to study history. I'm fascinated by ways it can be applied that would be either impractical or impossible to accomplish manually. What I've done here can be done (and has, obviously) without the magic of computers. But in the hands of a more skilled programmer than I, text mining offers up the real ability for both deep analysis and a degree of flexibility that goes beyond the typical scope of traditional methodology. When paired up with the massive digitization projects going on already that lowers the barrier to processing digital data (and, I fully admit, presents its own issues and problems), I think the tradeoff between the quality of results vs. time/effort is going to continue to shift in favor of text mining.

Larry H Cebula - September 14, 2009 @ 11:07 pm

Cameron: Thanks for taking my comments in the friendly spirit in which they are intended.

As a profession, we have been here before. In the 1960s the term Cliometrics was coined. Historians created punch cards based on census data and city directories and so on. It was going to REVOLUTIONIZE EVERYTHING. But nothing much ever came from it so far as I know. The one book title that pops up in my mind is Fogel and Engerman's Time on the Cross–a controversial book.

And yet–I am pretty sure that the application of digital technology is actually going to revolutionize everything–eventually. I want it to work.

Can you point me towards some historical text mining scholarship that has produced unique and compelling insights?

- **Ben Brumfield - September 15, 2009 @ 5:57 am**
  Larry, I think you make a fair point here. Text searching (not necessarily text mining) some parts of some scholarship, according to Patrick Leary's Googling the Victorians, but in most cases it's probably harder to figure out the questions to ask than to do the programming.

In my on project, for example, writing an analysis tool to look for correlation among subjects I'd already extracted was the matter of a single evening's hacking. But is it really that insightful to see that stripping tobacco occurs alongside clouds and rain? Or that mentions of the tenant farmer are common next to plowing? So far the most use I've gotten out of the tool has been in identifying unfamiliar names during the annotation process by looking for the context in which they're mentioned. Which is nice, but that's only happened *twice* in a few hundred pages. Web searches for unfamiliar names have worked just as often.

Despite those modest–even disappointing–results, I'm not sorry I built the tool, not least because it required such a modest effort. I think that perhaps we're moving beyond the model of large scale, resource-intensive text mining projects with unrealistic expectations to a model in which text mining is just another tool in the humanist's chest. Like a set of Allen wrenches: you may not need them very often, but they only cost a couple of bucks so you don't mind the expense.

Leave a Reply

Your email address will not be published / Required fields are marked *

Name*

Email*

Website

Post Comment

Notify me of follow-up comments by email.

Notify me of new posts by email.

October 19, 2009 by Cameron Blevins

Text Analysis of Martha Ballard's Diary (Part 3)

One of the most basic applications of text mining is simply counting words. I began by stripping out punctuation (in order to avoid differentiating **mend** and **mend.** as two separate words), put every word into lowercase, and then ignored a list of stop words (**the**, **and**, **for**, etc.). By writing a program to count occurrences of the 500 most common words, I could get a general (and more quantitative) sense for what general topics Martha Ballard wrote about in her diary. Unsurprisingly, her vocabulary usage followed a standard path of exponential decay: like most people, she utilized a relatively small number of words with extreme frequency. For example, the most common word (**mr**) occurred 10,050 times, while her 500th most common word (**relief**) occurred 67 times:

Because each word has information attached to it – specifically what date it was written – we can look at long-term patterns for a particular word's usage. However, looking at only raw word frequencies can be problematic. For example, if Ballard wrote the word **yarn** twice as often in 1801 as 1791, it could mean that she was doing a lot more knitting in her old age. But it could also mean that she was writing a lot more words in her diary overall. In order to address this issue,

for any word I was examining I made sure to normalize its frequency – first by dividing it by the total word count for that year, then by dividing it by the *average* usage of the word over the entire diary. This allowed me to visualize how a word's relative frequency changed from year to year.

In order to visualize the information, I settled on trying out sparklines: "small, intense, simple datawords" advocated by infographics guru Edward Tufte and meant to give a quick, somewhat qualitative snapshot of information. To test my method, I used a theme that Laurel Ulrich describes in *A Midwife's Tale*: land surveying. In particular, during the late 1790s Martha's husband Ephraim became heavily involved in surveying property. In the raw word count list, both **survey** and **surveying** appear in the top 500 words, so I combined the two and looked at how Martha's use of them in her diary changed over the years (1785-1812):

survey(ing)

Looking at the sparkline, we get a visual sense for when surveying played a larger role in Martha's diary – around the middle third, or roughly 1795-1805, which corresponds relatively well to Ulrich's description of Ephraim's surveying adventures. As a basis for comparison, the word **clear** appeared with numbing regularity (almost always in reference to the weather):

clear

Using word frequencies and sparklines, I could investigate and visualize other themes in the diary as well.

Religion

Out of the 500 most frequent words in the diary, only three of them relate directly to religion: **meeting** (#28), **worship** (#143), and **god** (#220).

meeting

worship

god

Meeting, which was used largely in a religious context (going to a church meeting), but also in a socio-political context (attending town meetings), had a relatively consistent rate of use, although it trended slightly upwards over time. **Worship** (which Martha largely used in the sense of "went to publick worship"), meanwhile, was more erratic and trended slightly downwards. Finally, and perhaps most interestingly, was Martha's use of the word **god**. Almost non-existent in the first third of her diary, it then occurred much more frequently, but also more erratically over the final two-thirds of the diary. Not only was it a relatively infrequent word overall (**flax**, **horse**, and **apples** occur more often), but its usage pattern suggests that Martha Ballard did not directly invoke a higher power on a personal level with any kind of regularity (at least in her diary). Instead, she was much more comfortable referring to the more socially and community-based activity of attending a religious service. While a qualitative close reading of the text would give a richer impression of Martha's spirituality, a quantitative approach demonstrates how little "real estate" she dedicates to religious themes in her diary.

Death

death

dead

funeral

expired

interd

Most of the words related to death show an erratic pattern. There are peaks and valleys across the years without much correlation between the different words, and the only word that appears with any kind of consistency is **interd** (interred). In this case, word frequency and sparklines are relatively weak as an analytical tool. They don't speak to any kind of coherent pattern, and at most they vaguely point towards additional questions for study – what causes the various extreme peaks in usage? Is there a common context with which Martha uses each of the words? Why was **interd** so much flatter than the others?

Family

In this final section, I'll offer up a small taste of how analyzing word frequency can reveal interpersonal relationships. I used the particular example of **Dolly** (Martha's youngest daughter):

dolly

The sparkline does a phenomenal job of driving home a drastic change in how Martha refers to her daughter. In a matter of a year or two in the mid 1790s, she goes from writing about **Dolly** frequently to almost never mentioning her. Why? Some quick detective work (or reading page 145 in *A Midwife's Tale*) shows that the plummet coincides almost perfectly with Dolly's marriage to a man named Barnabas Lambart in 1795. But why on earth would Martha go from mentioning **Dolly** all the time in her diary to going entire years without writing her name? Did Martha disapprove of her daughter's marriage? Was it a shotgun wedding?

The answer, while not so scandalous, is an interesting one nonetheless that text

analysis and visualization helps to elucidate. In short, Martha still writes about her daughter after 1795, but instead of referring to her as **Dolly**, she begins to refer to her as **Dagt Lambd** (Daughter Lambert). This is a fascinating shift, and one whose full significance might get lost by a traditional reading. A human poring over these detailed entries might get a vague impression that Martha has started calling her daughter something different, but the sparkline above drives home just how abrupt and dramatic that transformation really was. Martha, by and large, stopped calling her youngest daughter by her first name and instead adopted the new husband's proper name. Such a vivid symbolic shift opens up a window into an array of broader issues, including marriage patterns, familial relationships, and gender dynamics.

Conclusions

Counting word frequency is a somewhat blunt instrument that, if used carefully, can certainly yield meaningful results. In particular, utilizing sparklines to visualize individual word frequencies offers up two advantages for historical inquiry:

1. Coherently display general trends
2. Reveal outliers and anomalies

First, sparklines are a great way to get a quick impression of how a word's use changes over time. For example, we can see above that the frequency of the word **expired** steadily increases throughout the diary. While this can often simply reiterate suspected trends, it can ground these hunches in refreshingly hard data. By the end of the diary, a reader might have a general sense for how certain themes appear, but a text analysis can visualize meaningful patterns and augment a close reading of the text.

Second, sparklines can vividly reveal outliers. In the course of reading hundreds of thousands of words over the course of nearly 10,000 entries, it's quite easy to lose sight of the forest for the trees (to use a tired metaphor). Visualizing word frequencies allows historians to gain a broader perspective on a piece of the text, and they also act as signposts pointing the viewer towards a specific area for further investigation (such the red-flag-raising rupture in how frequently **Dolly** appears). Relatively basic word frequency by itself (such as what I've done here) does not necessarily explain anomalies, but it can do an impressive job of highlighting important ones.

#A Midwife's Tale *#Laurel Ulrich* *#Martha Ballard* *#Programming* *#Python* *#Text analysis*

Comments

Agnieszka Kielkiewicz-Janowiak - December 7, 2009 @ 5:19 am

Cameron,
I am really impressed by your work and dedication to getting an in-depth understanding Martha's story. I came across this wonderful text resource when I was researching New England women's private writings over 10 years ago. Working mostly from Poland, I found it invaluable to to be able to access this manuscript online (while I had to retrieve others from microfilm in a UMass library when I was there a short time). My special interest, as a sociolinguist, was language patterns and the occurrence of structures such as do-less negation, periphrastic do in declaratives, conjunctions (e.g. the almost obsolete "ere"), be and have with mutative intransitives, modal verbs, pronouns, etc. I was not just the word count I was after, but most importantly the context (broad and narrow). How I wish I had access to your expertise then! I published a book on the language of a few New England women in 2002. Thank you for renewing my interest in the text of martha's diary. And congratulations on your results!

- **Cameron Blevins - December 7, 2009 @ 9:27 am**
  Agnieszka,

  Thanks for the kind note. Your research sounds fascinating! I have close to zero background in sociolinguistics, but I can imagine the applications are pretty wide-ranging. I think one of the

fundamental challenges to my approach here is one of context – figuring out ambiguities or references without a human reader looking at each instance is tough. I'm hoping to devote some more time to exploring the diary in the next month or so, and will post whatever else I find here.

Thanks again!

-Cameron

Leave a Reply

Your email address will not be published / Required fields are marked *

Name*

Email*

Website

Post Comment

Notify me of follow-up comments by email.

Notify me of new posts by email.

April 1, 2010 by Cameron Blevins

Topic Modeling Martha Ballard's Diary

In *A Midwife's Tale*, Laurel Ulrich describes the challenge of analyzing Martha Ballard's exhaustive diary, which records daily entries over the course of 27 years: "The problem is not that the diary is trivial but that it introduces more stories than can be easily recovered and absorbed." (25) This fundamental challenge is the one I've tried to tackle by analyzing Ballard's diary using text mining. There are advantages and disadvantages to such an approach – computers are very good at counting the instances of the word "God," for instance, but less effective at recognizing that "the Author of all my Mercies" should be counted as well. The question remains, how does a reader (computer or human) recognize and conceptualize the recurrent themes that run through nearly 10,000 entries?

One answer lies in topic modeling, a method of computational linguistics that attempts to find words that frequently appear together within a text and then group them into clusters. I was introduced to topic modeling through a separate collaborative project that I've been working on under the direction of Matthew Jockers (who also recently topic-modeled posts from Day in the Life of Digital Humanities 2010). Matt, ever-generous and enthusiastic, helped me to install MALLET (Machine Learning for LanguagE ToolkiT), developed by Andrew McCallum at UMass as "a Java-based package for statistical natural language

processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text." MALLET allows you to feed in a series of text files, which the machine will then process and generate a user-specified number of word clusters it thinks are related topics. I don't pretend to have a firm grasp on the inner statistical/computational plumbing of how MALLET produces these topics, but in the case of Martha Ballard's diary, it worked. Beautifully.

With some tinkering, MALLET generated a list of thirty topics comprised of twenty words each, which I then labeled with a descriptive title. Below is a quick sample of what the program "thinks" are some of the topics in the diary:

- **MIDWIFERY:** birth deld safe morn receivd calld left cleverly pm labour fine reward arivd infant expected recd shee born patient
- **CHURCH:** meeting attended afternoon reverend worship foren mr famely performd vers attend public supper st service lecture discoarst administred supt
- **DEATH:** day yesterday informd morn years death ye hear expired expird weak dead las past heard days drowned departed evinn
- **GARDENING:** gardin sett worked clear beens corn warm planted matters cucumbers gatherd potatoes plants ou sowd door squash wed seeds
- **SHOPPING:** lb made brot bot tea butter sugar carried oz chees pork candles wheat store pr beef spirit churnd flower
- **ILLNESS:** unwell mr sick gave dr rainy easier care head neighbor feet relief made throat poorly takeing medisin ts stomach

When I first ran the topic modeler, I was floored. A human being would intuitively lump words like **attended**, **reverend**, and **worship** together based on their meanings. But MALLET is completely unconcerned with the meaning of a

word (which is fortunate, given the difficulty of teaching a computer that, in this text, **discoarst** actually means **discoursed**). Instead, the program is only concerned with how the words are *used* in the text, and specifically what words tend to be used similarly.

Besides a remarkably impressive ability to recognize cohesive topics, MALLET also allows us to track those topics across the text. With help from Matt and using the statistical package R, I generated a matrix with each row as a separate diary entry, each column as a separate topic, and each cell as a "score" signaling the relative presence of that topic. For instance, on November 28, 1795, Ballard attended the delivery of Timothy Page's wife. Consequently, MALLET's score for the **MIDWIFERY** topic jumps up significantly on that day. In essence, topic modeling accurately recognized, in a mere 55 words (many abbreviated into a jumbled shorthand), the dominant theme of that entry:

> "Clear and pleasant. I am a^{t} mr Pages, had another fitt of y^{e} Cramp, not So Severe as that y^{e} night pas^{t}. mrss Pages illness Came on a^{t} Evng and Shee was Deliverd a^{t} 11^{h} of a Son which waid 12 lb. I tarried all night She was Some faint a little while after Delivery."

The power of topic modeling really emerges when we examine thematic trends across the entire diary. As a simple barometer of its effectiveness, I used one of the generated topics that I labeled **COLD WEATHER,** which included words such as **cold**, **windy**, **chilly**, **snowy**, and **air**. When its entry scores are aggregated into months of the year, it shows exactly what one would expect over the course of a typical year:

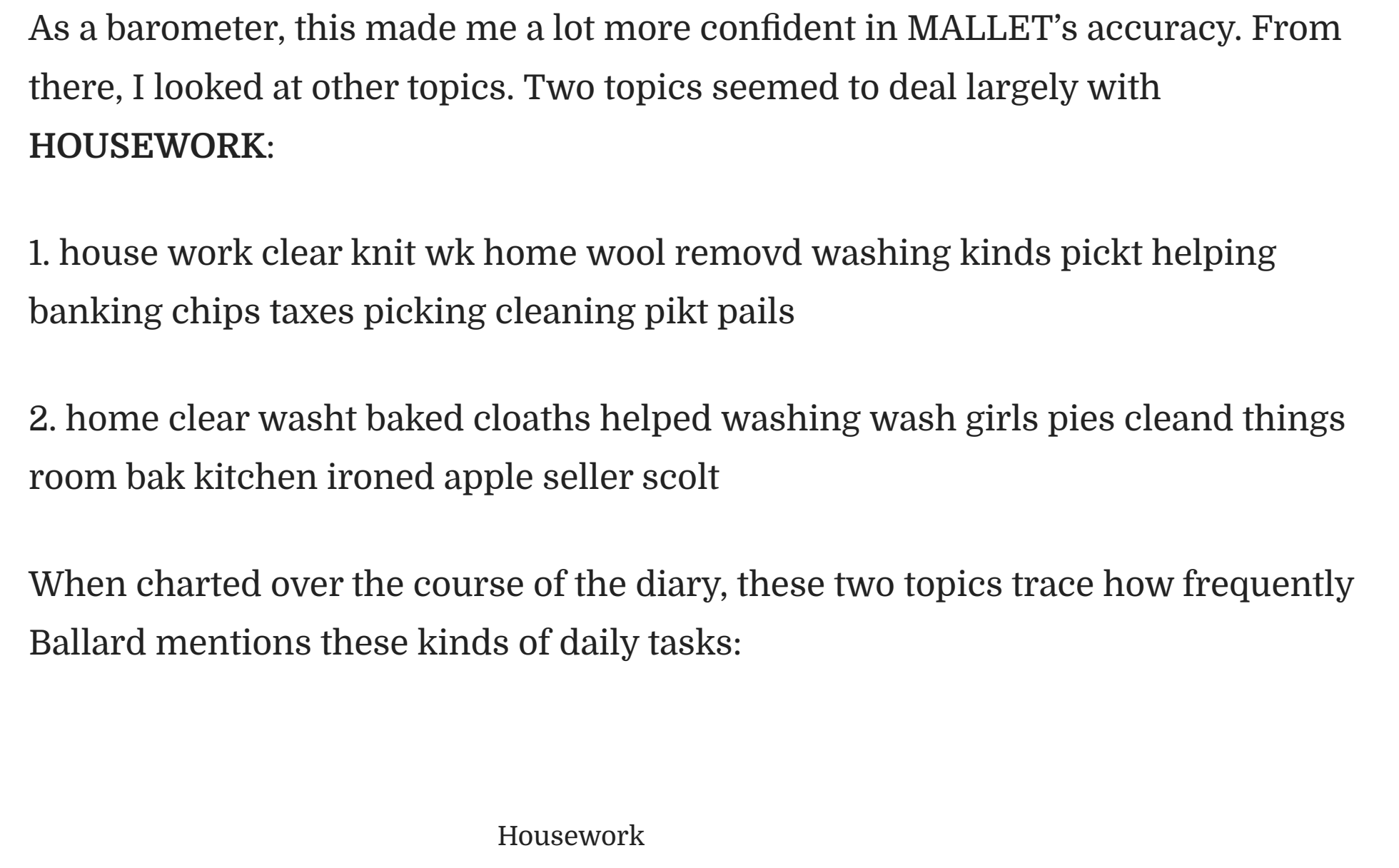
Cold Weather

As a barometer, this made me a lot more confident in MALLET's accuracy. From there, I looked at other topics. Two topics seemed to deal largely with **HOUSEWORK**:

1. house work clear knit wk home wool removd washing kinds pickt helping banking chips taxes picking cleaning pikt pails

2. home clear washt baked cloaths helped washing wash girls pies cleand things room bak kitchen ironed apple seller scolt

When charted over the course of the diary, these two topics trace how frequently Ballard mentions these kinds of daily tasks:



Both topics moved in tandem, with a high correlation coefficient of 0.83, and both steadily increased as she grew older (excepting a curious divergence in the last several years of the diary). This is somewhat counter-intuitive, as one would think the household responsibilities for an aging grandmother with a large family would *decrease* over time. Yet this pattern bolsters the argument made by Ulrich in *A Midwife's Tale*, in which she points out that the first half of the diary was "written when her family's productive power was at its height." (285) As her children married and moved into different households, and her own husband experienced mounting legal and financial troubles, her daily burdens around the house increased. Topic modeling allows us to quantify and visualize this pattern, a pattern not immediately visible to a human reader.

Even more significantly, topic modeling allows us a glimpse not only into Martha's tangible world (such as weather or housework topics), but also into her

abstract world. One topic in particular leaped out at me:

feel husband unwel warm feeble felt god great fatagud fatagued thro life time year dear rose famely bu good

The most descriptive label I could assign this topic would be **EMOTION** – a tricky and elusive concept for humans to analyze, much less computers. Yet MALLET did a largely impressive job in identifying when Ballard was discussing her emotional state. How does this topic appear over the course of the diary?

Emotion

Like the housework topic, there is a broad increase over time. In this chart, the sharp changes are quite revealing. In particular, we see Martha more than double her use of **EMOTION** words between 1803 and 1804. What exactly was going on in her life at this time? Quite a bit. Her husband was imprisoned for debt and her son was indicted by a grand jury for fraud, causing a cascade effect on Martha's own life – all of which Ulrich describes as "the family tumults of 1804-1805." (285) Little wonder that Ballard increasingly invoked "God" or felt "fatagued" during this period.

I am absolutely intrigued by the potential for topic modeling in historic source material. In many ways, it seems that Martha Ballard's diary is ideally suited for this kind of analysis. Short, content-driven entries that usually touch upon a limited number of topics appear to produce remarkably cohesive and accurate topics. In some cases (especially in the case of the **EMOTION** topic), MALLET did a better job of grouping words than a human reader. But the biggest advantage lies in its ability to extract unseen patterns in word usage. For instance, I would not have thought that the words "**informed**" or "**hear"** would cluster so strongly

into the **DEATH** topic. But they do, and not only that, they do so more strongly within that topic than the words **dead**, **expired**, or **departed**. This speaks volumes about the spread of information – in Martha Ballard's diary, death is largely written about in the context of news being disseminated through face-to-face interactions. When used in conjunction with traditional close reading of the diary and other forms of text mining (for instance, charting Ballard's social network), topic modeling offers a new and valuable way of interpreting the source material.

I'll end my post with a topic near and dear to Martha Ballard's heart: her garden. To a greater degree than any other topic, **GARDENING** words boast incredible thematic cohesion (gardin sett worked clear beens corn warm planted matters cucumbers gatherd potatoes plants ou sowd door squash wed seeds) and over the course of the diary's average year they also beautifully depict the fingerprint of Maine's seasonal cycles:

Gardening

Note: this post is part of an ongoing series detailing my work on text mining Martha Ballard's diary.

#Martha Ballard *#Programming* *#Text analysis* *#Text Mining* *#Topic Modeling*

Comments

Jason Boyd - April 1, 2010 @ 8:30 am

Fascinating. I work for Records of Early Drama (REED) — we publish collections of pre-1642 documents,

and I was very interested to see how effective MALLET was in dealing with a linguistically complex text like Martha Ballard's diary. Was the diary text you used marked up at all? Or was it a plain text file? Another question: although MALLET is unconcerned with word meanings, instead focussing on patterns of word usage, how does it overcome the problem of text that predates standardized spelling, punctuation, and grammar? Could it handle texts that were authored by numerous people over time, each of whom had their particular idiosyncrasies?

- **Cameron Blevins - April 1, 2010 @ 8:56 am**
  Jason,

  All good questions.

  1. The diary was not marked up at all. It was processed using Python into a basic list/array (with date, day of the week, text from the entry, etc.). From there I just exported the main text from each entry into ~10,000 separate .txt files, which MALLET could then treat as separate documents. Tracking them over time was a matter of naming the txt files by their date, such as 18070225.txt (2/25/1807).

  2. I was pleasantly shocked at how well MALLET handled the messiness of Ballard's shorthand style of writing. I think there were a few factors that contributed to this:

  – Stretched over 10,000 entries and 27 entries, the vagaries of different spellings tend to smooth out. Big data can overcome a lot of problems.

  – In a way, MALLET has an advantage in overcoming spelling variances. Provided the variances are somewhat consistent, it doesn't care whether the word is "delivd" or "delivered," all it knows is that particular string of characters tends to appear alongside "birth" words.

  3. MALLET can handle many different texts/authors – in fact, that's precisely what Matt Jockers has been doing. This has particular potential for clustering different authors together. The downside is that you tend to get "topics" that form based on unique words in an author's vocabulary. If you were to feed it contemporary British fiction, for instance, you'd probably get a topic of words like "Potter" "Hogwarts" and "Quidditch" – not particularly useful for analyzing trends your entire corpus. It all probably depends on just how variant the particular idiosyncrasies are from author to author.

  Hope this helps.

  -Cameron

erik steiner - April 2, 2010 @ 5:11 pm
Cameron,

This is awesome. I'm very intrigued by the possibility that this approach can be used to accurately model geographically varying patterns – such as climate. It would be very cool to track down actual weather data and correlate it with her references – or at least overlay it on your graphs. In theory, you could also reverse-geocode diaries (or newspapers) to determine based on their content where they were from. Since you know the locations of newspapers, it might be an interesting way to test this idea.

Also, I'm wondering about MALLET and the topics it defines – does it tell you how related two topics are to one another, and can you see this change over time? It would be interesting, for example, to see if Martha becomes has less EMOTION around DEATH as she gets older.

Great work. I look forward to more cool stuff from this.

-erik

- **Cameron Blevins - April 2, 2010 @ 11:32 pm**
  Erik,

  Thanks for the feedback. I really like the idea of reverse-geocoding, especially if you had a known-location training corpus for the program to work with.

  MALLET doesn't necessarily tell you how related two topics are to one another (at least I think, like I said I'm pretty shaky on how it works from a technical standpoint). But since I have all the temporal data associated with their "scores" for each entry, it's easy to do. I've actually played around a bit and set up a correlation matrix to see which topics move in tandem or apart. Mixed results so far, but it was interesting to see one topic that I was having trouble identifying move almost exactly opposite (coefficient of -0.9) with the COLD WEATHER topic over the course of a typical year. I still don't really know what the topic is (weakly associated with rainy weather?), but whatever it is seems to appear in the warmer months:

  cloudy afternoon rain home foren fore flax shower tn showers thunder af aft combd heavy turns misty dress pulld

  -Cameron

david blei - April 6, 2010 @ 10:01 am
this is fascinating.

re: geocoding. i work a lot on developing topic modeling tools. we recently developed a topic model that might account for location, by associating each document with a location and encoding which locations are adjacent to each other. (it's not exactly geocoding, but it kind of gets you there…)

we wrote about it in this paper, which is forthcoming from the annals of applied statistics:

http://www.cs.princeton.edu/~blei/papers/ChangBlei2010.pdf

the code is implemented in the "lda" R package. (in fact, this package lets you fit a number of types of topic models.)

best
dave

- **Cameron Blevins - April 6, 2010 @ 9:11 pm**
  Dave,

  Thanks for the comment! Although most of your paper was a bit over my non-quanty humanities head, it was interesting to see the intersection of topic modeling and geographic analysis. I'll also be sure to check out the LDA package, thanks for the suggestion.

  -Cameron

Lisa - April 17, 2010 @ 12:50 pm
Hi Cameron:

Thanks to you and Matt for introducing MALLET — I found your analysis of the product very interesting. I'm curious to know whether MALLET would also work for languages/scripts other than English? Say, Chinese?

By the way, the Archivist of the United States' most recent blog entry on the Library of Congress' acquisition of Twitter. He references Martha Ballard's Diary.
http://blogs.archives.gov/aotus/?p=172

Thanks again for a fascinating read.

- **Cameron Blevins - April 19, 2010 @ 9:31 am**
  Lisa,

  I'd be interested to see if it works on other languages, could have some fascinating potential there.

  Thanks for the link to the Archivist post, that was an interesting analogy between Ballard's diary entries as tweets.

  -Cameron

David Mimno - May 18, 2010 @ 7:20 am
Hi Cameron,

Thanks for using our MALLET topic modeling tools! This is exactly the type of research that got me

interested in statistical text mining.

Regarding irregular spellings: I've run this code on large early English collections, and it tends to find "clusters" of spelling variations, rather than smoothing over all variation and all time. For example you usually don't get 17th century spellings mixed with fully modern orthography. For a single-author corpus like this diary, it should work very well even with substantial variation.

On multiple languages: MALLET will support any language, although you may need to do some extra work creating "stoplists" of very common words and tokenizing the text (for example using the Stanford Chinese word segmenter). If you have documents aligned across multiple languages (such as wikipedia articles), MALLET also supports "polylingual" topic modeling: use the option –language-inputs instead of –input to learn topics in many languages simultaneously.

-David

- **Cameron Blevins - May 18, 2010 @ 8:03 pm**
  David,

  And thanks to you all at UMASS for building and maintaining such a great tool.

  I'm interested to hear about your experience with different corpora, especially ones that encompass several centuries. Do you think it's finding clusters of spelling variations because of the actual spelling patterns themselves, or their placement in the text? I think one reason MALLET seems to work so well on this is the fact that it's a single author, but I haven't had much experience with larger (and broader, or polylingual) corpora.

  Please send along my appreciation to the rest of the MALLET team.

  -Cameron

Steven - September 8, 2010 @ 11:18 pm
Hey,

One question anyone used MALLET on Social Media data specifically on blogs?

FM

Ron - October 4, 2011 @ 5:08 am
Any particular reason to use the one "l" spelling of "Topic Modeling"?

Datafiend - April 14, 2014 @ 4:31 am
Reblogged this on Austen, Morgan and Me and commented:
Detailed blog post exploring the use of MALLET to topic model a diary.

Choosing a platform for your project website

Learning HTML is no longer a requirement for building a website for your project. There are many platforms-- general-purpose platforms and ones tailored to specific kinds of projects-- that allow you to build much more sophisticated project sites than would be possible if you were building from scratch. When choosing a platform for your project website, the major factors to consider include functionality, familiarity, community, support, and cost.

FUNCTIONALITY

What do you want your project site to do? Are you developing an exhibit or collection of material, which needs to be displayed in a sequential order? Are you developing a directory, that you want to be browsable and searchable based on metadata you've entered (like "author", "publication date", "media used", etc.)? Do you want to use your site to transcribe content, or add annotations? Will users be able to create their own accounts, and will having an account provide them with additional access or unlock new tools on the site? What format(s) does your content take (text, audio, video, still images, downloadable files, etc)? Will your content be stored on the site itself, or is it coming from another hosting provider, like YouTube, a library website, or an institutional repository? How do you want to display your content-- in an image gallery, a timeline, a map, a list, or some other way?

These are just a handful of the considerations that should influence your decision about what platform to choose for your project. Functionality is the most important factor to address. While the other factors-- familiarity, community, support, and cost-- can help you choose

between multiple options that provide more or less the same level of functionality, choosing a platform that does not do what you need, because it's free or support is available for it is only a good idea as a stop-gap measure (e.g. to establish a URL and web presence in time for a conference or grant proposal) while you explore better options.

Ideally, you'd find a platform that does everything you want your project site to do, with minimal extra configuration, allowing you to focus on preparing and entering your data. This rarely happens, in part because what you want your site to do is often tied to the unique traits of your data itself, which a pre-made system isn't designed to accommodate without a little (or a lot) of work.

Some platforms, like Drupal, are extremely generic, and almost certainly won't do what you want out-of-the-box. To build a scholarly project site with Drupal, you have to add numerous "modules"-- or pieces of packaged-up functionality that someone has written code for. Drupal has a large international community of developers who write modules, and most modules can be installed and configured without you or your assistants ever having to look at the underlying code. Choosing a generic platform requires more time investment upfront, but leaves you with more flexibility later. For instance, if your project starts off with only text, but you later decided to incorporate video, it may be considerably easier to make that change if you've chosen a generic platform like Drupal (which some people use for textual content, others for video content, others for spreadsheet-like data) than a platform designed specifically for managing texts.

Other platforms make it very easy to build certain types of sites. Omeka is an example of a platform for sharing collections and exhibits. There are "add-ons" available for Omeka that extend its functionality, much like Drupal "modules", but they are intended to improve its collections and exhibits, not to fundamentally transform it into a different kind of platform, as add-ons for more generic platforms sometimes can (e.g. the BuddyPress "plugin", which turns WordPress from a blogging and generic content management platform into a social networking platform). Similarly, MediaWiki is a platform for building a wiki, and its "extensions" generally provide additional wiki functionality.

If you're not sure how your project may evolve, and all things are equal with regard to the other factors (like community, cost and support), you may be better off choosing a generic

platform, to keep your options open. If you have a clear sense of the limits of the project's scope, it may be better to choose a more specialized platform, if an appropriate platform exists.

FAMILIARITY

Platforms that started out having very different user interfaces are increasingly converging around certain design approaches and choices. All commonly-used platforms have (or can have, with the help of a module or plugin) a text authoring interface with WYSIWYG capabilities ("what you see is what you get" -- e.g. buttons you can click on to do things like make the font bold, or add a link, rather than making the user write HTML). Designs, or themes, that you can download and use for sites running any platform are increasingly adopting adaptive or responsive design techniques, which render the site differently depending on whether it's being viewed on a high-resolution laptop, a tablet, or a phone. This convergence makes it easier to make choices about the platform for your site without overly concerning yourself with what platforms are already being used by other popular sites in your field: chances are, you can make your site behave like other sites, even if it's running on a different platform. That said, the more you expect users to interact with your site-- be it through adding new content, or providing transcriptions or annotations, or engaging with other users-- the more important it is to minimize the learning curve required of your users. One of the easiest ways to do that is using platforms and plugins that they're already familiar with. For example, if you're building a scholarly network, and you know your users are MLA members, you may want to use the Commons In A Box package, which powers MLA Commons.

COMMUNITY

"Community" here refers to the group of people who are using the platform. Do other scholars in your field, or in related fields, use the platform that you are considering? A platform that's widely used by scholars may be a better choice than a platform whose major user base is small business owners, but if all the example sites you can find come from the sciences, where their data is considerably different than yours, you may want to make sure the platform meets your needs. Choosing a platform that's already being used by a community of humanities scholars may make it easier for you to ask questions, and get tips and advice on

how to deal with problems that arise, without having to translate your questions into language more easily understandable for technologists without a humanities background or scholars in another discipline.

SUPPORT

Who can help you develop a site using this platform, and what skills are required to do so? Most universities provide faculty with free access to some sort of web publishing platform, and offer training workshops and/or one-on-one consultation. Many universities also have a web development group with professional staff who may be available to consult or directly help you build your project site, at lower-than-market rates, but they may place restrictions on what platforms you can choose.

If support from a central or departmental IT group isn't an option (either because it's unavailable, or because the platforms they support are truly a bad fit for your project), there are ways of finding support on your own. There may be formal or informal meet-up groups around certain platforms, which provide an opportunity for people who are using the platform to exchange tips and suggestions, and you may be able to find someone in one of those groups who could do some freelance work. (The Berkeley Drupal Users' Group is one example.) You may be able to find a graduate or undergraduate student who can provide you with hands-on assistance, but it's important to know what you're looking for. Do you need someone to help you configure the platform by selecting, installing, and configuring a set of "modules" that are already available? This work requires a significantly less technical skill set than if you need someone to write new code to provide functionality that doesn't currently exist. If you need the person to write code, be sure you know what language(s) they will have to use-- it can vary depending on the platform, and the nature of what you need done. Depending on how elaborate your site design is, you might need to look for someone who has experience developing themes for your platform, which can be a very different skill set than writing modules. Your hosting choices may also be relevant here: are you hosting the site with a service that takes care of setting up the database and installing the platform, or do you need support from someone who can do that kind of work? Step-by-step installation guides can be found for all major platforms, but someone who's mostly comfortable configuring modules for the platform may not be comfortable working on the server level.

With the exception of writing code for modules, which does require specialized knowledge, building a site using most platforms is not beyond the capabilities of a curious humanities graduate student, given some time and opportunity to experiment and take advantage of the numerous how-to guides, books and video tutorials available online.However, some platforms may be more appealing to learn than others, especially for students considering alternative academic careers.

COST

The cost of a project website takes many forms-- hosting, configuration, ongoing maintenance, and the cost of developing new modules, if needed. Sometimes these costs are bundled together, for instance, if you're using proprietary software that's developed, hosted and maintained by a company. In most cases, though, you'll have to estimate these costs, which can vary wildly: inexpensive commercial hosting can cost around $100/year whereas deluxe packages where you have dedicated server resources can cost $100/month; undergraduates available through a research apprenticeship program may work for course credit, while professional web developers can charge $100+/hour. If you're using a freely-available open source platform, and you don't need to have new modules developed, your major costs will be site configuration and data entry. Finding the time to learn the platform well enough to do the configuration work yourself, and/or having a research assistant do that work, can cut costs considerably.

Because of the interplay of these various factors, it's hard to provide a general recommendation about what platform to use. Here are some commonly used platforms (all free and open source):

- Drupal - a general-purpose platform with a large international community of developers, including in higher ed and digital humanities. The Berkeley Drupal Users Group meets monthly on campus. A new working group of people who use Drupal for research will start meeting in spring semester 2014; contact Quinn (quinnd@berkeley.edu) for more information.
- MediaWiki - wiki platform used by Wikipedia and the Brueghel Family Research Website.
- Omeka, free/paid hosted version available at http://www.omeka.net/; designed for publishing collections/exhibits.
- Scalar is often used for multi-modal projects, and excels in multimedia annotation.

The Scalar site includes a few examples of projects that use it.

- WordPress - great for simple web publishing and blogging out of the box; lots of plugins are available to turn WordPress into a platform for text annotation (CommentPress), social networking (BuddyPress), etc.

If you'd like to discuss what might be a good fit for your project, please email digitalhumanities@berkeley.edu or use our contact form.

DIGITAL HUMANITIES AT BERKELEY

Site designed by **Agile Humanities** in association with Intelligent Machines.

1-1-2009

Digital Editions: Scholarly Tradition in an Avant-Garde Medium

Andrew Jewell
University of Nebraska at Lincoln, ajewell2@unl.edu

Digital Editions: Scholarly Tradition in an Avant-Garde Medium

Andrew Jewell

I come to you from the digital world. I am president of the *Digital* Americanists; I recently received a *Digital* Humanities Start-Up grant; I edit a *digital* archive on the life and work of Willa Cather; I am a faculty fellow at the Center for *Digital* Research in the Humanities; and I even have the word "*digital*" in my official job title: Assistant Professor of *Digital* Projects. I don't begin this way in order to impress you with my credentials, but as a confession: my current professional identity is absolutely entangled with the digital medium. That said, I want to confess something further: I am neck deep in the digital not because I have any particular interest in computers, but because our present—and future—academic environment is intertwined with this medium. The computer is a tool and a way to seize an opportunity to be what I really desire to be: an editor, a scholar, and teacher of literature.

In this way, I come from the digital world not really for idealistic reasons, but for circumstantial and pragmatic ones. In our current professional environment, there is a lot of energy and attention paid to the digital humanities and the dreamy new world it is ushering in, and it is becoming increasingly difficult to finance and publish large, sophisticated scholarly editions in print. Funding agencies are now demanding that editions be published in digital format, and the success of certain editorial projects in drawing in funds and attention—Ken Price's *Walt Whitman Archive*, for example—suggests that future developments in the field will likely require sophisticated engagement with computers.

Much of the rhetoric surrounding the new medium, however, is misleading, as it suggests the world as we know it is being *fundamentally* transformed. For example, a talk given by Brett Bobley, director of the NEH's Office of Digital Humanities, calls the presence of technology in the humanities "game-changing." The transformational rhetoric is important to the agendas of funding agencies and university administrators who need to convince constituents of their bold visions. And, to some degree, it is true: the digital medium does

indeed transform important elements of our scholarly work. However, it is also possible to see the trend toward digital humanities as a reclamation of scholarly traditions. G. Thomas Tanselle, in his insightful foreword to *Electronic Textual Editing*, writes:

> Even those engaged in textual criticism and scholarly editing have sometimes been swept along by the general euphoria and lost their sense of perspective. Their concerns, after all, are at the heart of the new developments, for what the computer offers . . . is a new way of producing and displaying visible texts. It can be of such great assistance to editors and other readers that they would be foolish not to make use of it and be excited about it. But when the excitement leads to the idea that the computer alters the ontology of texts and makes possible new kinds of reading and analysis, it has gone too far. The computer is a tool, and tools are facilitators; they may create strong breaks with the past in the methods for doing things, but they are at the service of an overriding continuity, for they do not change the issues that we have to cope with.

Tanselle's point has been borne out in my own educational and professional experiences: my work with digital editions has simultaneously forced me to learn new technologies and established traditions. Is has been an act of learning how to put a contemporary tool to the service of an established scholarly need. In fact, it was the digital humanities that introduced me to scholarly traditions that had no visibility in my undergraduate or graduate work in literary study. Until I worked applying XML markup to Walt Whitman's poetry manuscripts as a Graduate Research Assistant and engaged in debates about proper editorial policies, I had not been asked to confront elements of textual criticism: What is the role of authorial intention? What textual features are worthy of special editorial apparatus? What is the most effective form of annotation? How does one determine document order when leaves have become physically separated? Or, even more fundamentally: what is the most accurate transcription of this messy, handwritten document? The dominance of cultural studies and other theoretical models in the literary studies curriculum I encountered meant that work with texts and textual history was largely invisible. In fact, I'm embarrassed to say, I did not even know what a scholarly edition was until my graduate work was well under way.

Though my evidence is anecdotal, I believe that the excitement surrounding digital humanities has enabled a small surge in textual scholarship. At the University of Nebraska-Lincoln, where I work, one of the best-funded and most often-celebrated humanities initiatives is the Center for Digital Research in the Humanities. The institution expends significant resources to produce scholarly

works in a digital medium. Though these works of digital scholarship are widely varied, in most cases they involve some degree of documentary editing: transcription, markup, page scanning, proofreading, and more. The growth at UNL isn't unique, of course: digital humanities centers are popping up around the world in different forms, funding agencies are prioritizing digital work, and University presses are looking (sometimes boldly, sometimes not) to reclaim their sagging bottom lines and sense of purpose using digital technology. In that sense, the digital medium is creating an atmosphere in which more people are engaging with textual and documentary editing; or, to put it crassly, digital technology has helped people rediscover that textual work is really cool.

All of the labor required for digitizing has meant that significant numbers of undergraduates, graduate students, library staff members, and faculty members in a variety of departments are engaging in some aspect of documentary editing. Though it would go too far to claim that each person who encounters one of these projects gets a full education in the subject, it is true that hands-on work with texts, which necessitates some level of intellectual engagement with issues of textuality, is happening broadly, and with many, especially faculty and upper-level graduate students, it is happening deeply. The act of marking up a text in Text Encoding Initiative (TEI) conformant XML requires the encoder to decide what features of the text need markup and to provide an accurate transcription. In my interactions with students who are collaborating with me on my projects, we regularly converse about such matters as proper name regularization, placement of annotation references, and identification of structural markers in nineteenth-century newspapers. I can say with certainty—and there are failed grant applications to prove it—that digitization and the cultural cache that came with it made those conversations possible. Without the draw of the digital, my students and I would not be engaged with the same editorial issues. The enthusiasm engendered by the promise of new digital models of scholarship is what drew the students and resources to these projects. Tanselle counters this enthusiasm for digital technology with a crucial reminder of what it is we are doing when we engage with texts in the digital medium: "We should be enthusiastic about the electronic future, for it will be a great boon to all who are interested in texts; but we do not lay the best groundwork for it, or welcome it in the most constructive way, if we fail to think clearly about just what it will, and what it will not, change. Procedures and routines will be different; concepts and issues will not. . . . We will be spared some drudgery and inconvenience, but we still have to confront the same issues that editors have struggled with for twenty-five hundred years." Tanselle articulates an important point: the trend toward digital humanities is a *boon* for textual work, but it is not a fundamental remaking of it.

However, even if the fundamental intellectual issues are the same, the details are markedly different in the digital age. For an edition I'm working on,

the first complete, annotated edition of Willa Cather's extensive journalism, digital technology was not selected just to make it tenable in the current academic marketplace. Digital technology was selected because it made the edition better and more effective at communicating its content. These texts, for the most part, appeared once in Cather's lifetime, and that original publication exists only in the newspaper microfilm reels of the Nebraska State Historical Society. Additionally, these texts, though vibrant and highly readable to a modern audience, are choked with references to late nineteenth-century theater and popular culture, people and titles so well-known in 1894 that mere mention of the name was rhetorically adequate. With our digital edition, Kari Ronning and I can present the full texts of each of the 600 articles in an easily readable and searchable diplomatic transcription; we can provide a high quality page image of the original publication, which provides an authoritative image of the text and a glimpse into the fascinating context of the page; and we can provide thousands of annotations complete with images and, potentially, other media. The content of our edition of Cather's journalism could not exist in a print volume.

The edition of Willa Cather's journalism is only a part of the bigger digital project which I edit, the *Willa Cather Archive* (http://cather.unl.edu). This project is not exactly, or only, an edition. It is, more formally, what Carole Palmer calls "a new genre of scholarly production," a thematic research collection. Thematic research collections are, in Palmer's words, "digital aggregations of primary sources and related materials that support research on a theme" and are made because "[s]cholars have recognized that information technologies open up new possibilities for re-creating the basic resources of research and that computing tools can advance and transform work with those resources." It contains not just texts, but image galleries, interactive tools, and initiatives to organize communication among the community of Cather scholars. It is a project without a defined ending point that depends on collaborations with a wide range of people: undergraduates, graduate students, technical specialists, administrators, and scholars around the country. The thematic research collection is, in its most ambitious form, an attempt to digitally gather all the basic materials for one subject together in one place, to provide every reader, student, and scholar access to materials that traditionally have only been available to the privileged few that could afford to travel to archives around the world and carefully examine physically dispersed materials. Digitization can allow anyone with a web browser to see the documents only the elite have been able to see in the past.

This coexistence of a formal scholarly edition with other digitized materials under the same URL does perhaps blur for some the important distinction between "digitization" and "edition." The popularity of mass digitizing initiatives, from library-driven digital library projects to Google Books, have proliferated shabbily edited texts in electronic form, and this also suggests a possible threat to

the careful work of the editor. For example, textual scholar Wesley Raabe has tracked the way digital versions of Stowe's *Uncle Tom's Cabin* have transmitted inadequate versions of the texts, primarily by basing the transcribed, digital text on faulty reprint editions. And the digital versions have life beyond the screen, for the easy accessibility of digital editions appears to have made them the go-to texts for new print editions. As Raabe argues, "Print and digital traditions have become intermingled, and the status accorded to print editions in citation, when compared to the suspicion toward digital texts, is to misunderstand our contemporary textual condition" (Raabe 2008). Raabe's research provides an example of textual transmission concerns with a big text-digitization operation, one without particular concern for the specific content but instead interested in generating lots of electronic texts. The failure of mass digitization projects to provide excellent texts is unsurprising, and we understand that the motivation for the digitization—the "mass"—precludes rigorous copyediting.

But for other content-focused projects, the thematic research collections, the blur between digitization and editions is more complex and subtle. For many texts on the *Willa Cather Archive*, we make no claim to scholarly edition, nor do we even use the word "edition" to describe those materials. For other parts of the site, however, we are doing a full-on scholarly edition with full apparatus. This means that users are given different reading experiences for different texts: sometimes only a digital transcription is presented, more often users get a digital transcription combined with full-color page images of the original publication, and in one section users get the transcription, the page images, and extensive annotations.

This variety may trouble some, but the *Cather Archive*, though it is based largely on a collection of texts, does not consider itself at heart to be a big "scholarly edition." Instead, it *contains* such editions within a broader thematic research collection. It is meant to be a meaningful site for students and scholars studying Willa Cather, and the needs of those users—and the wide variety of multimedia materials available—means that, for some materials, a scholarly edition is required, but for other materials, it is more important that we provide access to forms not readily available (for example, our collection of Cather short fiction texts is made up predominantly of digital forms of her original periodical publications, complete with the accompanying illustrations which most readers of Cather have never encountered before.) I provide this description to reflect the way digital technology is allowing an edition to coexist with other materials not traditionally wedded so closely to it. Though to some the thematic research collection appears to be new world, in many ways this profusion of forms under one URL—images, sounds, video, interactive visual tools, and texts—is simply a multiformat extension of the drive behind documentary editions. The *Cather Archive*, though it may exist in different forms, is only trying to bring the pri-

mary materials important to its subject before as many people as it can in the most intellectually responsible and appropriate way possible.

In his opening paragraph of his essay on documentary editing in the *Electronic Textual Editing* volume, Bob Rosenberg is unequivocal about the connections between digital editions and their print forebears:

> The most important point to be made about any digital documentary edition is that the editors' fundamental intellectual work is unchanged. Editors must devote the profession's characteristic, meticulous attention to selection, transcription, and annotation if the resulting electronic publication is to deserve the respect given to modern microfilm and print publications. At the same time, it is abundantly clear that a digital edition presents opportunities well beyond the possibilities of film and paper.

I want to end today with some brief thoughts about what kinds of opportunities I can see with digitization, some of which will be entirely familiar, and others of which might be more unusual, but all of which I believe emerge out of the same concerns and desires that brought documentary editing into existence in the first place.

One of the most obvious benefits of digitization is the elimination of certain kinds of boundaries inherent in print volumes. In the digital environment, editors need not be so selective, but instead can contain all the texts they have the resources and moxie to produce, and they can present those texts as both searchable transcriptions and high-quality color images. In the presentation of texts, editors can choose multiple interfaces instead of just one: for example, if the text is encoded properly, one can alternate between a revision-ridden diplomatic transcription and a critical clear reading text with a click of a button. Or, one can allow users to browse edited documents chronologically or alphabetically or by any other arrangement that makes sense to the material being edited. The dynamism of the interface gives editors the chance to rid themselves of the tortured symbolic systems used in print to indicate various elements of the manuscript page and variants in different readings. Though rendering complex textual relationships is rarely straightforward, the digital environment's accessibility to color, animation, photographs, and space expands options considerably and allows us to dream of intuitive reading interfaces for our editions.

Once the texts are created, digital technology also allows readers to do more than just read them. Textual analysis gives users access to quantifiable data about the texts, information about word usage, phrase patterns, and grammatical choices. Willa Cather's readers can go to the *Cather Archive* and, thanks to Brian Pytlik Zillig's TokenX text analysis tool, gather unprecedented information

about the complete corpus of her fiction. They can see, for example, that she used the words "edit," "document," and "text" less than 20 times in her fiction, but used "book" or "books" hundreds of times (426 to be exact), or they can locate the most commonly used words and phrases used in sample texts. The value of these numbers will, of course, be determined by the value of the searches made and the interpretation of the numbers provided; the information does not replace interpretation, but gives the interpreters another piece of evidence to evaluate. One day, we hope to allow users to use increasingly sophisticated versions of this tool to track her language usage across time and across genres, to compare her language usage to her contemporaries, and to introduce part-of-speech analysis.

All of this, though, is simply an extension of an old motivating force: we want to give as many people as possible reliable and contextualized access to quality materials we consider important to the study of our subjects. Even the cutting-edge text analysis, though perhaps confounding for some modern literary scholars, would be recognizable to medieval monks who toiled on the first biblical concordance. In fact, the afternoon my colleague Brian Pylik Zillig showed me a recently generated list of all of the words Cather used in her fiction, I remarked, "Congratulations, Brian. You've just accomplished in a few minutes what some scholars used to take their entire careers to do." The tools we now use may be more complex and sophisticated than tools used in the past, but they are still at the service of the same basic scholarly challenges.

This paper was presented at the 2008 ADE Annual Meeting in Tucson, Arizona.

Works Cited

Bobley, Brett. "Why the Digital Humanities?" From a presentation given to the National Council on the Humanities. http://www.neh.gov/ODH/About/tabid/56/Default.aspx

Palmer, Carole L. "Thematic Research Collections." *Blackwell Companion to Digital Humanities.* Ed. Susan Schreibman, Ray Siemens and John Unsworth. Oxford: Blackwell, 2004. http://www.digitalhumanities.org/companion/

Raabe, Wesley. "Jewett's *Uncle Tom's Cabin*: A Case Study of Textual Transmission in the Digital Age." *The American Literature Scholar in the Digital Age.* Ed. Amy Earhart and Andrew Jewell. Ann Arbor: U of Michigan Press, forthcoming 2009.

Raabe, Wesley. Email to author. September 29, 2008.

Rosenberg, Bob. "Documentary Editing." *Electronic Textual Editing.* Ed. Lou Burnard, Katherine O'Brien O'Keeffe, and John Unsworth. New York: Modern Language Association, 2006. ["Preview" accessed on Text

Encoding Initiative website: http://www.tei-c.org/About/Archive_new/ETE/Preview /rosenberg.xml]

Tanselle, G. Thomas. "Foreword." *Electronic Textual Editing.* Ed. Lou Burnard, Katherine O'Brien O'Keeffe, and John Unsworth. New York: Modern Language Association, 2006. ["Preview" accessed on Text Encoding Initiative website: http://www.tei-c.org/About/Archive_new /ETE/Preview/tanselle.xml]

Guidelines for Editors of Scholarly Editions

Last revised 4 May 2022

1. Guidelines for Editors of Scholarly Editions1.1. Principles1.2. Sources and Orientations1.2.1. Considerations with Respect to Source Material1.2.2. The Editor's Theory of Text1.2.3. Medium (or Media) in Which the Edition Will Be Published2. Guiding Questions for Vetters of Scholarly Editions3. Glossary of Terms Used in the Guiding Questions4. Annotated Bibliography: Key Works in the Theory of Textual Editing

1. Guidelines for Editors of Scholarly Editions

1.1. PRINCIPLES

The scholarly edition's basic task is to present a reliable text: scholarly editions make clear what they promise and keep their promises. Reliability is established by

- accuracy
- adequacy
- appropriateness
- consistency
- explicitness

—accuracy with respect to texts, adequacy and appropriateness with respect to documenting editorial principles and practice, consistency and explicitness with respect to methods. The means by which these qualities are established will depend, to a considerable extent, on the materials being edited and the methodological orientation of the editor, but certain generalizations can be made:

- Many, indeed most, scholarly editions achieve reliability by including a general introduction —either historical or interpretive—as well as explanatory annotations to various words, passages, events, and historical figures.
- Scholarly editions generally include a statement, or series of statements, setting forth the history

of the text and its physical forms, explaining how the edition has been constructed or represented, giving the rationale for decisions concerning construction and representation. This statement also typically describes or reports the authoritative or significant texts and discusses the verbal composition of the text—its punctuation, capitalization, and spelling—as well as, where appropriate, the layout, graphic elements, and physical appearance of the source material. Statements concerning the history and composition of the text often take the form of a single textual essay, but it is also possible to present this information in a more distributed manner.

- A scholarly edition commonly includes appropriate textual apparatus or notes documenting alterations and variant readings of the text, including alterations by the author, intervening editors, or the editor of this edition.
- And finally, editors of scholarly editions establish and follow a proofreading plan that serves to ensure the accuracy of the materials presented.

1.2. SOURCES AND ORIENTATIONS

1.2.1. Considerations with Respect to Source Material

- *Is the date of the material known?* For example, in William Blake's *The Marriage of Heaven and Hell*, because the work itself bears no date, the date and its place in the author's oeuvre have to be inferred, and on such inferences other editorial decisions (decisions based, for example, on authorial intentions, which may vary over time) may depend. More generally, the location of a text in time and place may influence the editorial representation of a text.
- *Is there an author? La chanson de Roland*, for example, took a specific written form after a long life as a heroic poem or poems delivered orally from memory. Folktales, which may or may not originate with individual authors, are usually known to editors only in forms that have been shaped by transmission through communities of performers and listeners. W. B. Yeats and Georgiana Yeats claimed to have taken dictation from the spiritual world. Sacred texts are often attributed to divine authors or divinely inspired human authors.
- *Is the author known?* Authorship has been one of the most powerful and influential categories of textual criticism, where the "authority" of a text has often been determined by its convenient proximity to a known author writing in a specifiable time and space (traditionally, texts that come from an author's hand, such as an autographic manuscript, tend to have more authority in an edition than texts published after the author's death). When a text (for example, *Lazarillo de Tormes*) has no known author in the modern sense, or when authorship has been collaborative or communal, or when texts have taken shape over an extended period of time, editorial decisions must be based on other grounds.
- *Is there more than one author?* For example, Francis Beaumont and John Fletcher collaborated in writing over a dozen dramatic works between 1606 and 1616, such as *The Knight of the*

Burning Pestle; in addition to working together, these two writers also corrected and collaborated on texts with numerous other playwrights, including William Rowley, Philip Massinger, Thomas Middleton, and Ben Jonson, making it difficult, if not impossible, to assign authorship in some of these works to any one specific individual. Harriet Mill's role in the authorship of J. S. Mill's *Autobiography* might be labeled coauthorship; Theodore Dreiser sometimes revised his novels on the advice of a circle of family, friends, and associates. Max Perkins might be considered the coauthor of the novelists he edited as an employee of Scribner's—most notably Thomas Wolfe, whose published novels bear little resemblance to the manuscripts that Wolfe turned over to Perkins.

- *If there is an author (or authors), how far back in the process of authorship is source material available?* For example, there are no surviving manuscripts or working drafts for the majority of Daniel Defoe's more than 250 works, including his novels, such as *Moll Flanders* and *Robinson Crusoe*. The editor must rely instead on printed texts produced during Defoe's lifetime as the earliest sources.
- *Does the author play any other roles in producing the object being edited?* For example, Vladimir Nabokov translated his own early works from Russian into English, at a later point in his career; Blake printed and watercolored his illuminated books with the assistance of his wife, Catherine; Charles Dickens became his own publisher, first as an editor of *Bentley's Miscellany*, then as founder and editor of *Household Words* and *All the Year Round*.
- *How many other people are involved in producing the object being edited, and what are their roles?* For example, John Wilmot, the earl of Rochester, never published any of his works during his lifetime. Some of his poems were printed without his authority in songbooks and miscellanies, and they were widely circulated and preserved in manuscript copies. The subsequent posthumous editions gathered together many of these scattered pieces, but a modern editor must untangle the numerous variations found in the verses collected from these various manuscript and unauthorized printed versions. Another example would be the famously vexed case of James Joyce's *Ulysses*, drafted in longhand, typed by a typist, typeset by printers who spoke no English, and reset as many as five times, after Joyce's editing of page proofs.
- *Is it important, and is it feasible, to reproduce the material sources in facsimile as part of the edition?* A facsimile reproduction of an author's manuscript (or diary, or letters, or draft of an unpublished poem or novel) may make it easier to follow the process of composition than any translation of the manuscript into typographic form. For example, recent editors of Emily Dickinson have argued that something important is lost when Dickinson's "jottings" on scraps of paper are translated to the more familiar form of printed poems. In principle, it would seem always desirable to reproduce the source material for a scholarly edition in facsimile, but in print editions it is often impractical, and even in electronic editions it may be too expensive, or it may be impossible for lack of permission.

1.2.2. The Editor's Theory of Text

Editorial perspectives range broadly across a spectrum from an interest in authorial intention, to an interest in the process of production, to an interest in reception, and editors may select a given methodology for a variety of reasons. In very general terms, one could see copy-text, recensionist, and best-text editing as being driven by an interest in authorship—but best-text editing might also be driven by an interest in the process of production, along with "optimist," diplomatic, scribal, documentary, and social-text editing. Social-text editing might also be driven by an interest in reception—as "versioning" and variorum editing might be. And, of course, an editing practice that is primarily interested in authorship might very well be interested in production or reception or both—any good editor will be aware of the importance of all these things. However, when an editor has to choose what to attend to, what to represent, and how to represent it, there should be a consistent principle that helps in making those decisions. See the CSE's "Annotated Bibliography: Key Works in the Theory of Textual Editing," below, for further information on editorial methods and perspectives.

1.2.3. Medium (or Media) in Which the Edition Will Be Published

The decision to publish in print, electronically, or both will have an impact on a number of aspects of the edition, on its fortunes, and on the fortunes of its editor. Some questions an editor should consider in choosing the medium of publication:

- Is the source material itself manuscript, printed, electronic, or a combination of formats?
- What is the desired or potential audience for the work? Is there more than one audience? Will one medium reach the desired audience more effectively than another?
- What rights and permissions are required for publication, and do the terms differ by medium?
- What kind of apparatus can the edition have, and what kind should it have?
- Are there standard symbols or methods in a given medium for representing the typography, punctuation, or other textual features of the material being edited (Peirce's symbols, Shelley's punctuation, size-of-letter problems, spacing problems)?
- What is the importance of facsimile material, color reproductions, multiple versions, multiple states, interactive tools in this edition?
- Working with and from originals is of utmost importance; but some photographic, digitized reproductions make visible certain marks that have deteriorated and are no longer visible to the naked eye, even in the best light. If legibility has been enabled by the photographic or digitizing process, has that fact been explicitly noted to readers?
- How important is permanence or fixity? How can these qualities be attained?
- Alternatively, is there a possible benefit to openness and fluidity (for example, the certainty that new material will come to light)?
- Is there a publisher willing to publish in the medium you choose?

- How important is peer review (and if it is important, how will it be provided)?

2. Guiding Questions for Vetters of Scholarly Editions

Download the guiding questions.

Title vetted: _____Edited by: _____Date vetted: _____Vetter: _____

For each question listed below, the vetter should enter *Yes*, *No*, or *Not applicable* as appropriate. Vetter should also indicate whether additional comment on this point is made in the attached report.

		Y	N	N/A	See Report
I. Basic Materials, Procedures, and Conditions					
Materials					
1.0	Has the editor missed any essential primary or secondary materials?				
Stemma					
2.0	Has the editor accounted for the interrelations of all relevant texts?				
2.1	Have you tested the validity of the genealogy, stemma, or other account of the relevant texts against the collation data and included your findings in the report?				
Transcription					
3.0	Have all transcriptions been fully compared by the editor with the original documents, as distinct from a photocopy of those documents?				
3.1	If any transcriptions have not been fully compared with the originals, is there a statement in the edition alerting the user to that fact?				
3.2	Has someone other than the original transcriber carried out a thorough and complete check of each transcription, whether against the original or a photocopy of the original?				

3.3	Have you sampled the transcriptions for accuracy and included the results of that sampling in your report?				
Collation					
4.0	Have all potentially significant texts been collated?				
4.1	How many times have the collations been repeated by different people?				
4.2	Have you sampled the collations for accuracy and included the results of your sampling in your report?				
II. Textual Essay					
Principles and Methods					
5.0	If the edition under review is one in a series, have you examined textual essays and vetters' reports (if any) from earlier volumes?				
5.1	Does the textual essay provide a clear, convincing, and thorough statement of the editorial principles and practical methods used to produce this volume?				
5.2	Does it adequately survey all pertinent forms of the text, including an account of their provenance?				
Publication History and Physical Description					
6.0	Does it give an adequate history of composition and revision?				
6.1	Does it give an adequate history of publication?				
6.2	Does it give a physical description of the manuscripts or other pertinent materials (including electronic source materials, if any)?				
6.3	Are ways in which photographic or digital reproductions manipulate the text (sometimes leading to greater legibility) plainly described?				
6.4	Does it give a physical description of the specific copies used for collation?				
Copy-text					
7.0	Does the textual essay provide a convincing rationale for the choice of copy-text or base text or for the decision not to rely on either?				

7.1	Does it adequately acknowledge and describe alternative but rejected choices for the copy-text or base text?				
7.2	If there are forms of the text that precede the copy-text or base text, can they be recovered from the edited text and its apparatus?				
7.3	If not, is it practical, desirable, or necessary to make them recoverable?				
Changes to the Text					
8.0	Does the editor give an adequate account of changes to the text made by authors, scribes, compositors, et cetera?				
8.1	Are such changes to the text reported in detail as part of the textual apparatus?				
8.2	If such changes are recorded but the record will not be published, has the decision not to publish it been justified in the textual essay?				
Emendation					
9.0	Is the rationale for emendation of the copy-text or base text clear and convincing?				
9.1	Are all emendations of the copy-text or base text reported in detail or described by category when not reported in detail?				
9.2	Are the emendations of the copy-text or base text consistent with the stated rationale for emendation?				
9.3	Do the data from collation support the editor's assertion of authority for emendations drawn from the collated texts?				
9.4	If the author's customary usage (spelling, punctuation) is used as the basis for certain emendations, has an actual record of that usage been compiled from this text and collateral texts written by the author?				
9.5	Have you sampled the edited text and record of emendations for accuracy, and have you included the results in your report?				
9.6	Are emendations recorded clearly, avoiding idiosyncratic or ill-defined symbols?				
Illustrations and Typography					

10.0	Does the essay somewhere include an adequate rationale for reproducing, or not, the significant visual or graphic aspects of the copy-text or base text?				
10.1	Are all illustrations in the manuscript or the printed copy-text or base text reproduced in the edited text?				
10.2	If not, are they adequately described or represented by examples in the textual essay?				
10.3	Are the visual aspects of typography or handwriting either represented in the edited text or adequately described in the textual essay?				
10.4	If objects (such as bindings) or graphic elements (such as illustrations) are reproduced in the edition, are the standards for reproduction—sizing, color, and resolution—explicitly set forth in the textual essay?				
III. Apparatus and Extratextual Materials					
Nature of Collation					
11.0	Has a full historical collation been compiled, whether or not that collation is to be published?				
11.1	Is the rationale clear and convincing for publishing a selective historical collation (e.g., one that excludes variant accidentals)?				
11.2	Does the selective collation omit any category of variants you think should be included or include any you think should be excluded?				
11.3	Is the historical collation to be published accurate and consistent?				
Textual Notes					
12.0	Are the textual notes clear, adequate, and confined to textual matters?				
Ambiguous Textual Forms					
13.0	Have ambiguous hyphenated compounds (e.g., "water-wheel") in the copy-text or base text been emended to follow the author's known habits or some other declared standard?				
13.1	Have ambiguous stanza or section breaks in the copy-text or base text been consistently resolved by emendation?				

13.2	Are both kinds of emendation recorded in the textual apparatus to be published?				
13.3	For words divided at the end of a line in the edited text and stanzas or section breaks that fall at the end of a page in the edited text, can the reader tell how these ambiguous forms should be rendered when the text is quoted?				
Textual Apparatus					
14.0	Does the apparatus omit significant information?				
14.1	Can the history of composition and/or revision and/or the history of printing be studied by relying on the textual apparatus?				
14.2	Is the purpose of the different parts (or lists) in the apparatus clearly explained or made manifest?				
14.3	Is cross-referencing between the parts (or lists) clear?				
14.4	Is information anywhere needlessly repeated?				
14.5	Is the format of the apparatus adapted to the audience?				
14.6	Are the materials well organized?				
Accuracy of Extratextual Components					
15.0	Does the historical introduction dovetail smoothly with the textual essay?				
15.1	Has the editor quoted accurately from the edited text in the introduction and the textual essay?				
15.2	Has the editor verified references and quotations in the introduction and the textual essay?				
15.3	Has the editor checked the author's quotations and resolved the textual problems they present?				
15.4	Have you spot-checked to test the accuracy of quotation and reference in the introduction, textual essay, and text, and have you included the results of that spot-check in your report?				
Explanatory Notes					
16.0	Are the explanatory notes appropriate for this kind of edition—for example, in purpose, level of detail, and number?				

16.1	Is there a sound rationale for the explanatory notes, whether or not the rationale is to be made explicit anywhere in the published work?				
IV. Matters of Production					
State of Completion					
17.0	Did you see a final or near-final version of the edition or a substantial sample of it?				
17.1	If you did not see final or near-final copy, were you satisfied with the state of completion of the materials you did see?				
Permissions					
18.0	Has the editor obtained all necessary permissions—for example, to republish any materials protected by copyright?				
Publication Status					
19.0	If there is a publisher involved in producing the edition, has the publisher approved the content and format of the edition?				
19.1	Has the publisher approved the amount of time needed for proofreading?				
19.2	Has the publisher approved the requirements of the edition's design?				
19.3	Has the publisher approved cuing the back matter (textual apparatus and notes) to the text of the edition by page and line number (if this is a print edition) or by other unambiguous means (if this is an electronic edition)?				
19.4	Has the publisher approved the printer's or other production facility's copy requirements?				
Proofreading					
20.0	Has ultimate responsibility for maintaining accuracy throughout the production process been clearly assigned to one person?				
20.1	Are the proofreading methods sufficient to ensure a high level of accuracy in the published edition?				
20.2	If the editor supplies so-called camera-ready copy to the publisher, will it be proofread?				

20.3	How many proofreadings are scheduled?				
20.4	How many stages of proof are there?				
20.5	When a new stage of proof is read to verify changes or corrections, is adequate provision made for ensuring that all other parts of the text have not been corrupted?				
20.6	Is there a provision in place for collation or comparison of the first correct stage of proof against the production facility's final prepublication output (e.g., bluelines from a printer or text as rendered for final delivery in an electronic edition)?				
Editorial Rationale					
21.0	Does the project provide a sufficient justification of its technological choices, clarify the implications of these choices, and explain why these choices align with the editorial approach of the project?				
Preservation					
22.0	Does the edition have a preservation plan? For example, has an institution pledged long-term stewardship for the edition as part of its data preservation strategy? Has a copy of the edition and its multimedia elements, software, stylesheets, and documentation been deposited with a long-term digital object repository using file formats appropriate for preservation purposes?				
22.1	Has the edition included adequate documentation for reusing any data designated at shareable?				
22.2	Has a correction file or versioning system been set up and will it be maintained for tracking alterations to the edition after its initial publication?				
User Interface and Accessibility					
23.0	Does the edition provide a rationale for its accessibility standards? For example, does the edition follow the Web Content Accessibility Guidelines? Is it usable on multiple devices and by those in low bandwidth environments?				
23.1	Does the edition carry a clear statement of the appropriate reuse of its constituent elements, especially those protected by copyright, used by permission, or restricted by community protocols?				

23.2	Does the edition use a consistent framework? For example, if the text of the edition is encoded in an ISO standard grammar, such as XML, does it conform to relevant community guidelines? If the answer to the previous question is no, does the essay on technical methods provide a rationale for departing from community practice?				
23.3	Does the edition make available underlying data, distinct from presentation rendering? For example, if using markup, are the encoded files available to the reader for examination? If the underlying data are not available, is a rationale provided for restricting access to it?				
23.4	If the edition includes electronic files, are those files encoded in an open, nonproprietary format (e.g., TEI XML rather than *Microsoft Word* or *WordPerfect*)?				
23.5	Does the edition include help documentation that identifies the features of the user interface and explains how to use them?				
Documentation and Metadata					
24.0	Does the project document its digitization processes and standards (if applicable)?				
24.1	Does the edition employ relevant standards for its technical, descriptive, and administrative metadata?				
24.2	If the edition employs metadata that does not rely on an external standard or combines elements from multiple standards, is the data model documented?				
24.3	If any software has been uniquely developed for this edition, is source code for that software available and documented? If not, has a rationale for restricting access or a timeline for publishing the source code been provided?				

3. Glossary of Terms Used in the Guiding Questions

The glossary was drafted by Robert Hirst and subsequently revised and expanded by the committee.

accidentals: A collective term invented by W. W. Greg and now widely used to mean the punctuation, spelling, word division, paragraphing, and indications of emphasis in a given text—things "affecting mainly its formal presentation," as he put it ("The Rationale of Copy-Text," *Studies*

in Bibliography 3 [1950–51]: 21). Greg distinguished between the accidentals of a text and its words, or substantives (q.v.). Accidentals and substantives are conceptually important for Greg's rationale of copy-text, which assumes that authors are more proprietary about their words than about their accidentals, while typesetters and other agents of textual transmission (copyists, typists, proofreaders, copyeditors) are the reverse. For this reason, at least for an edition aimed at preserving the author's accidentals as well as substantives, the rationale for choosing a copy-text is first and foremost that, of the available texts, it is the most faithful to the author's accidentals and contains the fewest changes to them by other hands. It is therefore often the first or earliest text in a line of descent, but any author who carefully revised the accidentals (say, in the second edition) might oblige an editor to choose that text rather than an earlier one.

authority: A property attributed to texts, or variants between texts, in order to indicate that they embody an author’s active intention, at a given point in time, to choose a particular arrangement of words and punctuation. Authority therefore always derives from the author, even when *author* is defined and understood as coauthor, collaborator, or a collective (like the vorticists). Where the author is unknown or uncertain, authority will need to be argued. It is even possible to invert the usual pattern and assign authority to agents who produce variants commonly regarded as unauthoritative, such as typesetters, proofreaders, or reprint publishers—though one hesitates to call such agents "author." However defined, the author produces texts or variants that have authority. Some reprints may be said to have "no authority" because the author had no role in producing them. On the other hand, texts that were set from copy revised by the author are said to contain "new authority," meaning that some of their variants arose from the author's revision. The authority of a holograph manuscript is usually greater than any typesetting of it, but the manuscript's authority at any given point may be superseded if the typesetting incorporates authorial changes—a case of "divided authority."

base text: The text chosen by an editor to compare with other texts of the same work in order to record textual variation among them. Its selection can be to some extent arbitrary, or it can be selected because it is (among the available texts) simply the most complete. Unlike a copy-text (q.v.), it is not assigned any presumptive authority and may not even be used to construct a critical text, serving instead only as an anchor or base to record textual variants.

collation: Comparison. A collation is either the record of the substantive and accidental differences between two or more texts or the act of comparing two or more texts for the purpose of documenting their differences.

copy-text: The specific arrangement of words and punctuation that an editor designates as the basis for the edited text and from which the editor departs only where deeming emendation necessary. Under W. W. Greg's rationale the copy-text also has a presumptive authority in its accidentals (that is, the editor will default to them wherever variant accidentals are "indifferent" —meaning not persuasively authorial or nonauthorial). But *copy-text* may also designate texts for

which no later variants are possible or anticipated. It is now commonplace to designate a manuscript letter that was actually sent as a copy-text for a personal letter. In such cases, emendations of the copy-text would normally consist not of the author's subsequent revisions but solely of elements in the original manuscript that the editor could not, or elected not to, represent in the transcription. Contrary to certain common misconceptions, *copy-text* does not mean the copy an editor or author sends to the printer, and it need not represent the "author's final intention." Indeed it is more likely to be the author's first draft than the author's final printed revision of a text. Its selection is based on the editor's judgment that the authority of its accidentals is on the whole superior to other possible texts that could be chosen for copy-text.

digital object repository: A means of storing, retrieving, and administering complex collections of digital objects. If the repository is to meet the needs of scholarly editions, it should have a secure institutional basis (like a university research library), and it should have a commitment to long-term preservation, migration, and access. For an example, see http://fedorarepository.org/.

DTD (document type definition): The set of rules that specifies how the SGML or XML grammar will be applied in a particular document instance.

emendations: Editorial changes in the copy-text or base text. These changes may be made to correct errors, to resolve ambiguous readings, or to incorporate an author's later revisions as found in printed editions or other sources, such as lists of errata, assuming for the moment that the editorial goal is to recover the author's textual intentions. Different editorial goals might well call for emendations of some other kind, but they would all still be editorial changes to the copy-text or base text and would under normal circumstances be reported as part of the editor's accounting of the handling of available evidence.

end-of-line hyphens: Hyphens in a word that fall at the end of a line in a manuscript or in typeset material. End-of-line hyphens may sometimes be ambiguous. They may be either (a) signs of syllabic division used to split a word in two for easier justification of a line of type (or to fit it on the end of one and beginning of the next manuscript line) or (b) signs that a compound word is to be spelled with a hyphen. A word like *water-wheel* or *Jack-o-lantern* if broken after a hyphen at the end of a line might be ambiguous—that is, it is unclear whether the word is intended to be spelled with or without the hyphen. For any source text these ambiguous hyphens require judgment as to how the word was intended to be spelled, and such ambiguities would ordinarily be resolved in the way other ambiguous readings in a copy-text are resolved—by editorial choice, recorded as an emendation (change) in the copy-text. In the text as finally edited and printed, if hyphenation of certain words falls at the end of a line and is therefore ambiguous, the editor should likewise resolve this ambiguity for the reader.

explanatory notes: Notes devoted to explaining what something means or why it is present, rather than textual notes, which are devoted to explaining why the text at a certain point reads in the way

it does and not in some other way.

historical collation: A record of variants for a given text over some defined number of editions (e.g., from the first through the seventh editions) or some period of time (e.g., from different impressions of the same edition made between 1884 and 1891). The purpose of historical collations is to put before the reader as complete a record as possible of all variants among a group of texts from which the editor has had to choose. In the past, but only to save space, historical collations have tended to omit variant accidentals and confine themselves to a record of variant substantives.

ISO: The short name for the International Organization for Standardization, a worldwide federation of national standards bodies from more than 140 countries, one from each country. ISO is a nongovernmental organization established in 1947. The mission of ISO is to promote the development of standardization and related activities in the world with a view to facilitating the international exchange of goods and services and to developing cooperation in the spheres of intellectual, scientific, technological, and economic activity. See www.iso.org.

JPEG (Joint Photographic Experts Group), or JPG: An open, nonproprietary ISO standard (official name ITU-T T.81 | ISO/IEC 10918-1) for the storage of raster images. For more information, see www.jpeg.org.

machine collation: Collation by means of a Hinman Collator or other mechanical or optical device, allowing very slight differences between states of the same typesetting to be located visually, without the need for a traditional point-by-point comparison of one text against the other. Machine collation is only possible between different states of the same typesetting.

modernizing: Changing the spelling or punctuation of a text to bring these into conformity with modern standards, as distinct from the standards at the time of first composition or publication.

METS (Metadata Encoding and Transmission Standard): A standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium. The standard is maintained in the Network Development and MARC (machine-readable cataloging) Standards Office of the Library of Congress and is being developed as an initiative of the Digital Library Federation. For more information, see www.loc.gov/standards/mets.

MPEG (Moving Picture Experts Group): The nickname given to a family of international standards used for coding audiovisual information in a digital compressed format. The MPEG family of standards includes MPEG-1, MPEG-2, and MPEG-4, formally known as ISO/IEC-11172, ISO/IEC-13818, and ISO/IEC-14496. Established in 1988, the MPEG working group (formally known as ISO/IEC JTC 1/SC 29/WG 11) is part of JTC1, the Joint ISO/IEC Technical Committee on Information Technology. For more information, see https://mpeg.chiariglione.org/.

PNG (portable network graphics): An extensible file format for the lossless, portable, well-

compressed storage of raster images. The PNG specification is on a standards track under the purview of ISO/IEC JTC 1 SC 24 and is expected to be released eventually as ISO/IEC international standard 15948. See www.libpng.org.

raster image: An image stored and shown in terms of points, each one based on a set number of bytes that define its color. Arranged in a grid of pixels on a monitor, the points represent the tones, colors, and lines of the image. Common raster formats include TIFF and JPEG. Sometimes called bitmapped images, raster graphics are often contrasted to vector graphics, which represent images by such geometrical elements as curved lines and polygons rather than points in a grid. Vector graphics are typically used in programs for drawing and computer-aided design (CAD).

rendering process: The application of rules to transform content from storage format (e.g., TEI XML) to delivery format (e.g., XHTML), for the purpose of display in a Web browser. A vetter usually encounters these rules embodied in XSL stylesheets, but they could take other forms as well (PHP, CSS, etc.).

schema: A means for defining the structure, content, and semantics of XML documents. For more information, see www.w3.org/XML/Schema.

SGML (standard generalized markup language): A grammar for text encoding, defined in ISO 8879. For more information, see xml.coverpages.org/sgml.html.

silent emendations: Editorial changes to the copy-text that are not recorded, item by item, as they occur but are only described somewhere in the textual essay as a general category of change and are thus made "silently," without explicit notice of each and every change.

stemma: A schematic diagram representing the genealogical relation of known texts (including lost exemplars) of a given work, showing which text or texts any given later text was copied from, usually with the overall purpose of reconstructing an early, lost exemplar by choosing readings from later extant texts, based in part on their relative distance from the lost source. A stemma may also be used simply to show graphically how any given text was copied or reprinted over time, even if the goal is not to recover an early, lost exemplar.

substantives: W. W. Greg's collective term for the words of a given text—"the significant . . . readings of the text, those namely that affect the author's meaning or the essence of his expression," as distinct from its accidentals ("The Rationale of Copy-Text," *Studies in Bibliography* 3 [1950–51]: 21). Under Greg's rationale for copy-text, the authority for substantives could be separate and distinct from the authority for the accidentals, thus permitting an editor to adopt changes in wording from later texts, even though maintaining the accidentals of an earlier one virtually unchanged.

tag library: A document that lists all the tags, or elements, available in a DTD, with a brief description of the intended use of each, a list of its attributes, and statements identifying elements

within which this element can occur and which elements it can contain. See www.loc.gov/ead/tglib/index.html for an example.

textual notes: Notes devoted specifically to discussing cruxes or particular difficulties in establishing how the text should read at any given point. Compare "explanatory notes."

user interface: In an electronic edition, the on-screen presentation of content, including navigational methods, menus of options, and any other feature of the edition that invites user interaction or responds to it.

variants: Textual differences between two or more texts. These would include differences in wording, spelling, word division, paragraphing, emphasis, and other minor but still meaning-bearing elements, such as some kinds of indention and spacing.

XML (extensible markup language): A simplified subset of SGML (q.v.), developed by the World Wide Web Consortium. For a gentle introduction to XML, see www.tei-c.org/P4X/SG.html

XSL (extensible stylesheet language): A language for expressing stylesheets. An XSL stylesheet specifies the presentation of a class of XML documents (e.g., TEI documents) by describing how an instance of the class is transformed into an XML document that uses the specified formatting vocabulary (e.g., HTML). For more information, see www.w3.org/Style/XSL.

4. Annotated Bibliography: Key Works in the Theory of Textual Editing

Download the annotated bibliography.

The bibliography was drafted by Dirk Van Hulle and subsequently revised and expanded by the committee. For a more extensive compilation of works on the topic, see William Baker and Kenneth Womack, Twentieth-Century Bibliography and Textual Criticism: An Annotated Bibliography *(Westport: Greenwood, 2000).*

Bédier, Joseph. "La tradition manuscrite du *Lai de l'ombre*: Réflexions sur l'art d'éditer les anciens textes." *Romania* 54 (1928): 161–96, 321–56.

> Bédier advocates best-text conservatism and rejects the subjectivity of Karl Lachmann's method (see Maas), which emphasizes the lost authorial text, resulting remarkably often in two-branch stemmata. Instead, Bédier focuses on manuscripts and scribes, reducing the role of editorial judgment.

Biasi, Pierre-Marc de. "What Is a Literary Draft? Toward a Functional Typology of Genetic

Documentation." *Drafts*. Ed. Michel Contat, Denis Hollier, and Jacques Neefs. Spec. issue of *Yale French Studies* 89 (1996): 26–58.

In a continuous effort to present manuscript analysis and *critique génétique* as a scientific approach to literature, Biasi designs a typology of genetic documentation, starting from the *bon à tirer* ("all set for printing") moment as the dividing line between the *texte* and what precedes it, the so-called *avant-texte*.

Blecua, Alberto. "Defending Neolachmannianism: On the *Palacio* Manuscript of *La Celestina*." *Variants*. Ed. Peter Robinson and H. T. M. Van Vliet. Turnhout: Brepols, 2002. 113–33.

A clear position statement by the author of the noteworthy Spanish *Manual de critica textual* (1983) in defense of the neo-Lachmannian method. Blecua argues that stemmatic analysis is superior to the methods based on material bibliography and that only the construction of a stemma can detect the presence of contaminated texts.

Bornstein, George, and Ralph G. Williams, eds. *Palimpsest: Editorial Theory in the Humanities.* Ann Arbor: U of Michigan P, 1993.

On the assumption that texts are not as stable or fixed as we tend to think they are, these essays examine the palimpsestic quality of texts, emphasizing the contingencies both of their historical circumstances of production and of their reconstruction in the present. They mark a theoretical period of transition, shifting the focus from product to process in editorial theory and practice.

Bowers, Fredson. "Some Principles for Scholarly Editions of Nineteenth-Century American Authors." *Studies in Bibliography* 17 (1964): 223–28.

Concise and systematic elaboration of W. W. Greg's theories, arguing that "when an author's manuscript is preserved," this document rather than the first edition has paramount authority and should serve as copy-text. Bowers's principles for the application of analytic bibliography in an eclectic method of editing have been most influential in Anglo-American scholarly editing.

Bryant, John. *The Fluid Text: A Theory of Revision and Editing for Book and Screen*. Ann Arbor: U

of Michigan P, 2002.

Bryant draws attention to textual fluidity, which results from processes of revision. In terms of scholarly editing, this implies a method of representation that does not obviate but rather emphasizes moments of textual instability. Although the examples are mostly taken from Melville's works, the ideas are generally applicable to other writings.

Burnard, Lou, Katherine O'Brien O'Keeffe, and John Unsworth, eds. *Electronic Textual Editing*. New York: MLA, 2006.

The guidelines of the Committee on Scholarly Editions and the Text Encoding Initiative frame a collection of essays on both practical and theoretical issues in electronic textual editing, ranging from levels of transcription to the preservation of electronic editions. The collection's goal is to encourage careful work in the production of digital editions and to facilitate its evaluation so that scholars can receive professional credit for the transmission of cultural heritage from print to electronic media.

Bustarret, Claire. "Paper Evidence in the Interpretation of the Creative Process in Literary Manuscripts." *L'Esprit Créateur* 41.1 (2001): 16–28.

Bustarret analyzes the valuable clues paper analysis offers for understanding the complex interaction between the phases of writing and editing. She contends that paper is "not to be considered any more as a passive surface receiving the creative work, but as a tool in the creative process," and offers fascinating case studies of the writings of Proust, Duchamp, Roussel, and others.

Cohen, Philip, ed. *Devils and Angels: Textual Editing and Literary Theory.* Charlottesville: UP of Virginia, 1991.

The "increasingly theoretical self-consciousness" characterizing textual criticism and scholarly editing marks an impasse, indicative of a paradigm shift. Assumptions that have been self-evident for several decades are rethought in eight stimulating essays and three responses.

Deegan, Marilyn, and Kathryn Sutherland, eds. *Text Editing, Print and the Digital World*. Farnham: Ashgate, 2009.

Taking stock of recent trends in digital humanities and scholarly editing, this collection of essays clearheadedly assesses the state of the discipline. Instead of loudly announcing paradigm shifts, the editors allow divergent voices to examine how existing approaches are evolving and responding to new editorial challenges, both in theory and in practice.

Deppman, Jed, Daniel Ferrer, and Michael Groden, eds. *Genetic Criticism: Texts and Avant-Textes*. Philadelphia: U of Pennsylvania P, 2004.

This representative collection of eleven essays by French critics (such as Louis Hay, Pierre-Marc de Biasi, Almuth Grésillon, and Jean Bellemin-Noël) introduces genetic criticism (*critique génétique*) to an Anglo-American audience, distinguishing this critical mode from textual criticism. Apart from the elucidating general introduction, each of the essays is preceded by an informative introduction and bibliography.

Eggert, Paul. *Securing the Past: Conservation in Art, Architecture and Literature*. Cambridge: Cambridge UP, 2009.

This methodological study approaches textual scholarship as a form of preservation that is comparable to architectural conservation and painting restoration. The book provides a lucid survey of the history of textual scholarship and a theoretically informed definition of *the work* as a regulative principle in terms of a negative dialectic between changing signifying aspects and equally changing physical states.

———. "Textual Product or Textual Process: Procedures and Assumptions of Critical Editing." *Editing in Australia*. Ed. Eggert. Canberra: U Coll. ADFA, 1990. 19–40.

Starting from a comparison with new techniques of x-raying paintings, Eggert proposes a valuable ideal for a critical edition that allows the reader to study both the writing process and the finished product.

Ferrer, Daniel. "Production, Invention, and Reproduction: Genetic vs. Textual Criticism." *Reimagining Textuality: Textual Studies in the Late Age of Print*. Ed. Elizabeth Bergmann Loizeaux and Neil Fraistat. Madison: U of Wisconsin P, 2002. 48–59.

Ferrer defines the difference between genetic and textual criticism on the basis of their respective foci on invention and repetition. He pleads for a hypertextual presentation as the best way to do justice to the diverse aspects of the writing process.

Finneran, Richard J., ed. *The Literary Text in the Digital Age*. Ann Arbor: U of Michigan P, 1996. Editorial Theory and Lit. Criticism.

The availability of digital technology coincides with a fundamental paradigm shift in textual theory, away from the idea of a "definitive edition." Fifteen contributions reflect on the shift toward an enhanced attention to nonverbal elements and the integrity of discrete versions.

Fiormonte, Domenico. *Scrittura e filologia nell'era digitale*. Turin: Boringhieri, 2003.

Taking Italian *filologia* as his frame of reference, Fiormonte includes advances in various fields of research and different national contexts to expound his view on the theoretical implications of electronic editing and digital philology or "postphilology." In the appendix, "Risorse digitali per la filologia," Cinzia Pusceddu discusses a number of existing electronic editions and useful tools for textual criticism and digital philology.

Gabler, Hans Walter. "The Synchrony and Diachrony of Texts: Practice and Theory of the Critical Edition of James Joyce's *Ulysses*." *Text* 1 (1981): 305–26.

The work's "total text," comprising all its authorial textual states, is conceived as a diachronous structure that correlates different synchronous structures. A published text is only one such synchronous structure and not necessarily a privileged one.

Gabler, Hans Walter, George Bornstein, and Gillian Borland Pierce, eds. *Contemporary German Editorial Theory*. Ann Arbor: U of Michigan P, 1995.

With its representative choice of position statements, this thorough introduction to major trends in German editorial theory in the second half of the twentieth century marks the relatively recent efforts to establish contact between German and Anglo-American editorial traditions.

Gaskell, Philip. *From Writer to Reader: Studies in Editorial Method.* Oxford: Clarendon, 1978.

In 1972, as *A New Introduction to Bibliography* was replacing R. B. McKerrow's manual, Gaskell had already criticized W. W. Greg's copy-text theory, arguing that authors often expect their publishers to correct accidentals. *From Writer to Reader* zooms in on the act of publication and the supposed acceptance of the textual modifications this may involve.

Greetham, David C., ed. *Scholarly Editing: A Guide to Research.* New York: MLA, 1995.

The most comprehensive survey of current scholarly editing of various kinds of literatures, both historically and geographically, with elucidating contributions by textual scholars from different traditions.

———. *Textual Scholarship: An Introduction.* New York: Garland, 1992.

An impressive survey of various textual approaches: finding, making, describing, evaluating, reading, criticizing, and finally editing the text—namely, biblio-, paleo-, and typography; textual criticism; and scholarly editing. The book contains an extensive bibliography, organized by discipline.

Greg, W. W. "The Rationale of Copy-Text." *Studies in Bibliography* 3 (1950-51): 19–36.

This pivotal essay has had an unparalleled influence on Anglo-American scholarly editing in the twentieth century. Greg proposes a distinction between substantive readings (which change the meaning of the text) and accidentals (spelling, punctuation, etc.). He pleads for more editorial judgment and eclectic editing, against "the fallacy of the 'best text'" and "the tyranny of the copy-text," contending that the copy-text should be followed only so far as accidentals are concerned and that it does not govern in the matter of substantive readings.

Grésillon, Almuth. *Eléments de critique génétique: Lire les manuscrits modernes.* Paris: PUF, 1994.

An introduction to textual genetics or *critique génétique*, which was developed in the 1970s and became a major field of research in France. In spite of correspondences with textual criticism, it sees itself as a form of literary criticism, giving primacy to interpretation over editing.

Groden, Michael. "Contemporary Textual and Literary Theory." *Representing Modernist Texts: Editing as Interpretation.* Ed. George Bornstein. Ann Arbor: U of Michigan P, 1991. 259–86.

An important plea for more contact between textual and literary theorists, by the general editor of the *James Joyce Archive* facsimile edition of Joyce's works.

Hay, Louis. "Passé et avenir de l'édition génétique: Quelques réflexions d'un usager." *Cahier de textologie* 2 (1988): 5–22. Trans. as "Genetic Editing, Past and Future: A Few Reflections of a User." Trans. J. M. Luccioni and Hans Walter Gabler. *Text* 3 (1987): 117–33.

Genetic editing, presenting the reader with a "work in progress," is a new trend, but it revives an old tradition. The founder of the Institute for Modern Texts and Manuscripts (ITEM-CNRS), in Paris, points out that editing has always reflected the main ideological and cultural concerns of its day.

Kirschenbaum, Matthew G. *Mechanisms: New Media and the Forensic Imagination*. Cambridge: MIT P, 2008.

Just as the discipline of textual studies considers the physical traces of a writing process to examine variants, this study applies computer forensics to examine three important works of new media and electronic literature, paying attention to the specificity of multiple versions, storage devices, systems, and platforms.

Maas, Paul. *Textkritik*. Leipzig: Teubner, 1927. Vol. 2 of *Einleitung in die Altertumswissenschaft*. Trans. as *Textual Criticism*. Trans. Barbara Flower. Oxford: Clarendon, 1958.

One of Karl Lachmann's main disciples, Maas systematizes Lachmannian stemmatics, requiring thorough scrutiny of witnesses (*recensio*) before the emendation of errors and corruptions (*emendatio*, often involving a third step of divination or *divinatio*).

Martens, Gunter, and Hans Zeller, eds. *Texte und Varianten: Probleme ihrer Edition und Interpretation*. Munich: Beck, 1971.

An epoch-making collection of German essays with important contributions by, among others, Zeller (pairing "record" and "interpretation," allowing readers to verify the editor's decisions), Siegfried Scheibe (on fundamental principles for historical-critical editing), and Martens (on textual dynamics and editing). The collection's central statement is that the apparatus, not the reading text, constitutes the core of scholarly editions.

McGann, Jerome J. *Critique of Modern Textual Criticism*. 1983. Charlottesville: UP of Virginia, 1992.

Textual criticism does not have to be restricted to authorial changes but may also include the study of posthumous changes by publishers or other agents. McGann sees the text as a social construct and draws attention to the cooperation involved in the production of literary works.

———. "The Rationale of Hypertext." *Text* 9 (1996): 11–32. Rpt. in *Electronic Text: Investigations in Method and Theory.* Ed. Kathryn Sutherland. Oxford: Clarendon, 1997. 19–46. Rpt. in *Radiant Textuality: Literature after the World Wide Web.* New York: Palgrave, 2001. 53–74.

Conceived in an expressly revisionist relation to W. W. Greg's rationale, McGann's ambitious essay presents the book as a machine of knowledge and evaluates the advantages of hyperediting and hypermedia over editions in codex form. As the earliest hypertextual structure, the library organization illustrates the theoretical design of a "decentered text."

———. *The Textual Condition*. Princeton: Princeton UP, 1991.

McGann makes several valuable and innovative suggestions, from the idea of a "continuous production text" to a clear distinction between a text's bibliographic and linguistic codes (in the important essay "What Is Critical Editing?").

McKenzie, D. F. *Bibliography and the Sociology of Texts: The Panizzi Lectures, 1985.* London: British Lib., 1986.

McKenzie extends the scope of traditional bibliography to a broader sociology of the text, including video games, movies, and even landscapes. This perspective has been a major stimulus to the advancement of the sociological orientation in scholarly editing.

McKerrow, R. B. *An Introduction to Bibliography for Literary Students.* Oxford: Oxford UP, 1927.

McKerrow's manual of "new bibliography" reflects the early-twentieth-century editorial method that made extensive use of analytic bibliography. The author of *Prolegomena for the Oxford Shakespeare* was rather averse to the idea of emending the copy-text from other sources.

Mikhailov, Andreï, and Daniel Ferrer, eds. *La textologie russe: Anthologie*. Paris: CNRS, 2007.

Anthology of articles ranging from the early days of Russian formalism to recent criticism. Whereas French genetic criticism focuses on the destabilizing tendency of comparing different versions and prefers to separate the study of the writing process from the editorial impulse to provide a stable text, Russian textology (a term coined by Boris Tomashevsky in 1928) is concerned both with authors' creative processes and with the way their works are presented to the public.

Modiano, Raimonda, Leroy F. Searle, and Peter Shillingsburg, eds. *Voice, Text, Hypertext: Emerging Practices in Textual Studies.* Seattle: U of Washington P, 2004.

A collection of essays arguing that texts are not only "documents" or "material objects" but also "cultural events." Drawing from classical Roman and Indian to modern European traditions, the contributors reveal that "to study a text is to study a culture." Additionally, the essays suggest the role of textual scholarship in cultural studies and critical theory.

Nutt-Kofoth, Rüdiger. *Dokumente zur Geschichte der neugermanistischen Edition*. Tübingen: Niemeyer, 2005.

Collection of thirty-five important theoretical essays that have shaped the German editorial tradition, including texts by Karl Lachmann, Jacob Grimm, Karl Goedeke, Gerog Witkowski, and Reinhold Backmann. The book offers a historical survey of the discipline from the mid-eighteenth century until the publication of the important volume *Texte und Varianten* (1971), which marks the start of contemporary German editorial theory.

Nutt-Kofoth, Rüdiger, and Bodo Plachta, eds. *Editionen zu deutschsprachigen Autoren als Spiegel der Editionsgeschichte*. Tübingen: Niemeyer, 2005.

The essays in this collection focus on the development in editorial approaches to the work of twenty important German-speaking authors (in alphabetical order from Brecht to Trakl), plus one survey article on electronic editions by Fotis Jannidis. This second volume in the series Bausteine zur Geschichte der Edition is part of the editors' initiative to produce a history of scholarly editing.

Nutt-Kofoth, Rüdiger, Bodo Plachta, H. T. M. Van Vliet, and Hermann Zwerschina, eds. *Text und Edition: Positionen und Perspektiven*. Berlin: Schmidt, 2000.

As a younger generation's counterpart of *Texte und Varianten* (see Martens and Zeller), this state of the art of current scholarly editing in Germany also includes interesting survey articles on Anglo-American scholarly editing (e.g., Peter Shillingsburg) and "genetic criticism and philology" (Geert Lernout; trans. in *Text* 14 [2002]: 53-75).

Parker, Hershel. *Flawed Texts and Verbal Icons: Literary Authority in American Fiction.* Evanston: Northwestern UP, 1984.

Starting from analyses of revisions by Herman Melville, Mark Twain, Stephen Crane, and Norman Mailer, Parker pleads for more attention to textual composition and the development of (sometimes self-contradictory) authorial intentions, which an institutionalized editorial method is often unable to represent.

Pasquali, Giorgio. *Storia della tradizione e critica del testo.* Florence: Le Monnier, 1934.

Pasquali criticizes some of the basic Lachmannian principles and proposes to take the history of the witnesses and the scribes into account. The current emphasis on textual tradition in Italian philology is to a large extent his legacy.

Pizer, Donald. "Self-Censorship and Textual Editing." *Textual Criticism and Literary Interpretation.* Ed. Jerome J. McGann. Chicago: U of Chicago P, 1985. 144–61.

Pizer emphasizes the social aspects of texts, arguing that even when authors personally change their texts under external pressure, it may be more important to present the reader with the censored versions because of their social resonance.

Reiman, Donald H. "'Versioning': The Presentation of Multiple Texts." *Romantic Texts and Contexts*. Columbia: U of Missouri P, 1987. 167–80.

Reiman suggests "versioning" (or multiversional representation) as an alternative to "editing." The main purpose of this textual approach is to offer readers and critics the opportunity to figure out for themselves how the work evolved.

Robinson, Peter M. W. "The One Text and the Many Texts." *Making Texts for the Next Century*. Spec. issue of *Literary and Linguistic Computing* 15.1 (2000): 5–14.

Robinson, in answer to the question, "Is there a text in these variants?," which he asked in a previous essay, argues that a scholarly edition is more than merely presenting an archive of variants. The aim of the editor should be to offer a useful tool to allow readers to make the connection between variation and meaning. A critically edited text (presented along with "the many texts") is the best means to that end.

Schreibman, Susan, Ray Siemens, and John Unsworth, eds. *A Companion to Digital Humanities*. Oxford: Blackwell, 2004.

This collection of thirty-seven essays consolidates its broad, authoritative coverage of the emerging field of humanities computing in four sections: history; principles; applications; and production, dissemination, and archiving. Topics range from computer basics and digital textual editing to speculative computing, project design, and preservation.

Shillingsburg, Peter L. *From Gutenberg to* Google*: Electronic Representations of Literary Texts*. Cambridge: Cambridge UP, 2006.

The book's hypothesis is that the electronic representation of print literature will significantly alter our understanding of textuality. Shillingsburg's "script act theory," a synthesis of theories on written literary texts developed in separate fields, is the basis for a proposal of an electronic infrastructure for script acts, as well as for a negotiation of conflicting objectives in different editorial traditions.

———. *Resisting Texts: Authority and Submission in Constructions of Meaning*. Ann Arbor: U of Michigan P, 1997.

The editor's main task, Shillingsburg argues, is to relate the work to the documents and to take responsibility for the integrity of the agency of texts, which is a responsibility to both the author *and* the social contract. Shillingsburg designs a map with four major forms of textual concern, placing the physical documents at the center of textual and literary theory.

———. *Scholarly Editing in the Computer Age: Theory and Practice.* 3rd ed. Ann Arbor: U of Michigan P, 1996.

An indispensable introduction to practical procedures and controversial issues in editorial theory, offering clear definitions in matters of textual ontology and a survey of different orientations in scholarly editing.

Stillinger, Jack. *Multiple Authorship and the Myth of the Author in Criticism and Textual Theory*. New York: Oxford UP, 1991.

Stillinger pleads for a broader conception of authorship to include collaboration as an inherent aspect of creation. Case studies include John Stuart Mill and his wife, John Keats and his helpers, and William Wordsworth revising earlier versions of his texts.

Tanselle, G. Thomas. "The Editorial Problem of Final Authorial Intention." *Studies in Bibliography* 29 (1976): 167–211.

Authors' revisions do not automatically reflect their final intentions. In the case of *Typee*, Herman Melville was responsible for the changes in the second edition, but they represent his "acquiescence" rather than his intention, according to Tanselle, who is well aware that a reader does not have access to an author's mind and who advises editors to always take the context into account.

———. *A Rationale of Textual Criticism.* Philadelphia: U of Pennsylvania P, 1989.

In his profound analysis of the ontology of texts, Tanselle makes a clear distinction between *work* and *text*. A work is an entity that exists in no single historical document. Scholarly editing entails, just like any act of reading, the effort to discover the work that "lies behind" the text(s) one is presented with.

Théorie: État des lieux. Spec. issue of *Genesis: Revue internationale de critique génétique* 30 (2010): 1–300.

The thirtieth volume of the journal *Genesis* contains an inventory of different theories in textual studies, including Anglo-American, Italian, French, and German approaches, and explores common ground between genetic criticism and neighboring disciplines, such as literary history, sociocriticism, and digital humanities.

Thorpe, James. *Principles of Textual Criticism*. San Marino: Huntington Lib., 1972.

As an early critic of the principles advocated by W. W. Greg and Fredson Bowers, Thorpe argues that specific compositional peculiarities and contingencies tend to be left out of consideration.

Timpanaro, Sebastiano. *La genesi del metodo del Lachmann.* 1963. Padua: Liviana, 1985.

The genealogical study of manuscript transmission originated in New Testament criticism toward the end of the eighteenth century. By reexamining Joseph Bédier's criticism regarding two-branch stemmata, Timpanaro does not so much aim to correct them but to understand how they came into being.

Tisseron, Serge. “All Writing Is Drawing: The Spatial Development of the Manuscript.” *Yale French Studies* 84 (1994): 29–42.

Tisseron reasserts the value of the active, fluid processes of the “poetics of writing” as opposed to the “poetics of the text.” His essay explores the inscriptive process from the perspective of “the original spatial play which the hand stages,” noting especially connections between the somatics of writing and the process of thinking.

Vanhoutte, Edward. “Electronic Textual Editing: Prose Fiction and Modern Manuscripts: Limitations and Possibilities of Text-Encoding for Electronic Editions.” *Text Encoding Initiative*. TEI, n.d. Web. 22 Feb. 2011.

Vanhoutte offers a concise yet complex definition of the electronic edition and its principal aims. In addition to discussing editorial principles and markup, the author proposes a methodology that “might help us in combining the study of ‘what cannot be observed’ in very observable markup.’” Most significantly, Vanhoutte presents a way of “getting time back in manuscripts.”

Van Hulle, Dirk. *Textual Awareness: A Genetic Study of Late Manuscripts by Joyce, Proust, and Mann*. Ann Arbor: U of Michigan P, 2004.

The first part of the book gives a concise but thorough overview of the three editorial traditions (*Editionswissenschaft*, *édition critique* and *critique génétique*, and textual criticism and scholarly editing). The second part of the book is a genetic analysis of three major works of world literature: James Joyce's *Finnegans Wake*, Marcel Proust's *À la recherche du temps perdu*, and Thomas Mann's *Doktor Faustus*.

Zeller, Hans. "A New Approach to the Critical Constitution of Literary Texts." *Studies in Bibliography* 28 (1975): 231–63.

> In his evaluation of Anglo-American copy-text theory from a structuralist point of view, Zeller contrasts the practice of editing an "eclectic (contaminated) text" with German editorial methods, showing crucial differences with respect to the notions of "authority," "authorial intention," and "version."

Does the project provide a sufficient justification of its technological choices, clarify the implications of these choices, and explain why these choices align with the editorial approach of the project?

Modern Language Association MLA

MLA Statement on the Scholarly Edition in the Digital Age

MLA Committee on Scholarly Editions

Web publication, May 2016

MLA Statement on the Scholarly Edition in the Digital Age

Executive Summary

IN AN era of mass data, both the macro and micro scales of scholarly editions are being reimagined. Today, the scholarly edition can provide a single perspective on a text archive that supports large-scale textual research. In this sense, the scholarly edition, providing clear documentary evidence of the relations and contexts of primary materials, allows forms of analysis and engagement beyond those of its editorial intention, supporting further scholarship.

Digital modalities open up important opportunities for alternative uses of scholarly editions. First, they allow the data in an edition to be used as the basis for other editions, as transcriptions that can be compared using collation tools, as a contribution to a digital repository, and as part of a text corpus that might support quite different types of analysis. Second, digital modalities make it possible to support features such as user annotation, commentary, citation, and the creation of additional layers of editorial information. Third, digital modalities allow edition interfaces to serve as environments for manipulation and exploration of the edition's textual space, so that the user can occupy the role of a contingent editor.

At its inception and in its early documents, the Committee on Scholarly Editions (CSE) adopted a fairly specific definition of the kinds of editions it would cultivate and endorse. More recent CSE discussions have emphasized the need to broaden the scope of the CSE's attention to include different editorial modalities, highlighting the need for a set of standards of excellence that can generalize well across different types of editions. In its statement, the committee sets forth an initial set of minimal conditions that mark an edition as a scholarly edition today and identifies further conditions that apply specifically to a digital scholarly edition, including, but not limited to, the following:

- it must note its technological choices and be aware of their implications, ideally using technologies appropriate to the goals of the edition, in recognition of the fact that technologies and methods are interrelated in that no technical decisions are innocent of methodological implications and vice versa;
- it should be created and presented in ways ensuring the greatest chance of longevity—addressing this challenge involves infrastructural, financial, and data representation issues (such as the use of widely accepted, open standards);
- it should readily respond to the challenge of maintaining the scholarly ability to be referenced in view of the ways that interfaces change over time; and
- where possible, it should attend to possibilities of sampling, reuse, and remix, supporting approaches to the formation and curation of the edition such as reconstructing and documenting instances of texts and textual change over time, like algorithmic construction and reconstruction (with possible extensibility, including external data); in doing so, it should attempt to balance considerations for intellectual property and labor with the goals of achieving open access and reusability.

MLA Statement on the Scholarly Edition in the Digital Age

Preamble and Statement of Purpose

THIS statement is intended as a tool for thinking through a set of pressing questions for the MLA's Committee on Scholarly Editions (CSE) and as a contextualized expression of our current responses to those questions. These questions at bottom amount to "What is a (digital) scholarly edition?" and "How can the CSE, through its practices and guidelines, encourage excellence in (digital) scholarly editing?" These questions are of course not new, but the committee has not yet addressed them directly and formally in the context of digital editorial practice. This statement is an attempt to do so.

The audience for this document is threefold. First, it is intended for the present and future members of the CSE, for whom it will hopefully serve as a record of current ideas and rationales and help in anchoring policy or explaining decisions to later committee members. Second, it is intended for scholarly editors who are interested in using the CSE's guidelines and would like to know more about their intellectual background. Third, it is intended for a wider audience that includes those who are interested in scholarly editing and editions and are curious about the evolution of the committee's thinking.

The main focus of this document is definitional: we examine in turn the three major elements of the "digital scholarly edition" and explore their significance for the ways editions are read, used, and evaluated. In particular we consider a set of crucial features that we take to be fundamental to scholarly editing: transparency, accuracy, appropriateness of method, clear and responsible documentation, and the exercise of critical judgment in representing a full account of the textual situation at stake. We conclude by reflecting on the CSE's own responsibilities in the light of these definitions. However, it's worth posing at the outset a more fundamental question: What is the point of scholarly editions, as we currently understand them, in an era of mass data? One way to answer this question (anticipating some of the discussion further on in the document) is to note that a key trend in scholarly editing itself is toward the creation of an edition as a single perspective on a much-larger-scale text archive. The edition of Goethe's *Faust* being developed at the University of Würzburg, for instance, includes a documentary archive of all *Faust*-related materials by Goethe and a critical edition that draws its data from that archive. Even standing alone, these archives constitute a resource that supports large-scale textual research; if aggregated through a mechanism like *HathiTrust* or *TAPAS*, they constitute a body of material that is much larger and could be used for broader cultural analysis. The scholarly edition, in other words, is in some cases being rethought in a way that involves both the micro and the macro scales.

It is also important to recognize that although some kinds of large-scale cultural research can be conducted on informationally undifferentiated resources like *Google Books*, more nuanced research (and more powerful scholarly argumentation) requires data that are more representationally detailed. For instance, scholars studying the changing revision habits of generations of American novelists would need access to data that capture revision as an explicit informational component. The preparation of this kind of data—whether we call it scholarly editing or something else—draws on the same levels of expertise and care, and the same kind of attention to the specificities of texts, as the traditional scholarly edition, albeit applied toward new ends; this is a particularly fertile area for future collaboration among those in scholarly editing, bibliography, information studies, and the digital humanities more broadly construed.

Finally, along with the trend toward scale we are now also seeing a concomitant acknowledgment of the interdependence of micro- and macroanalysis, and we are seeing increased emphasis on approaches that enable scholars to move effectively between the two. The scholarly edition as we see it emerging here is well adapted to both kinds of work: it offers a detailed account of the data scholars need in order to make sense of a specific textual landscape, but it does so in a way that is formalized and programmatic and hence can support computational analysis at any scale. Through the use of standards like the TEI Guidelines, editions can also be studied in groups (though clearly this study requires careful coordination of efforts to make the data commensurable across editions).

Issues

We proceed by first considering how to define the central terms (*edition*, *scholarly*, and *digital*) for purposes of this work and then considering the minimal qualities a "digital scholarly edition" ought to have, how we can enable these qualities, and what further research questions arise that might further the field.

Edition

Our definition of an edition begins with the idea that all editions are mediations of some kind: they are a medium through which we encounter some text or document and through which we can study it. In this sense an edition is a re-presentation, a representational apparatus, and as such it carries the responsibility not only to achieve that mediation but also to explain it: to make the apparatus visible and accessible to criticism.

To unpack this further: an edition is a systematic account of a text guided by a specific theory of what such an account should be (e.g., one that is concerned with the genesis of a literary work, one that is concerned with the social ecology of the text, one that is concerned with contextualizing a single unpublished manuscript document). An edition is thus also a model, in the sense that it serves as an analytic surrogate for the textual landscape it describes, one that can be manipulated and queried to yield insight into its details. Although this definition sounds as if it might apply chiefly to digital editions, in fact it is also true of print: there the manipulation

in question may happen through the creation of multiple views of the same data (e.g., indexes, bibliographies, concordances). In print these are necessarily represented as distinct entities, but they constitute the same kind of approach to the data.

There are a few other aspects of a potential edition that are more controversial. If we consider the facsimile edition as a kind of limit case—one that takes to one logical conclusion the idea of putting curated textual materials before the reader within a framework that permits analysis and interpretation—then we need to consider what kinds and degrees of curation constitute editing for purposes of our definition. Among other things this thought experiment encourages us to consider the role of the edition as a *model* of a textual space that makes its contents tractable to analysis rather than as an aggregation that minimizes its mediation of those contents. Another issue is that of comprehensiveness and whether an edition needs to represent and curate all extant texts and documents. While comprehensiveness is desirable to the extent that it puts the reader in possession of a maximum amount of relevant information, there are certainly legitimate editorial situations where a focus on a single document, or on a limited subset of available documents, may be appropriate. In addition, in the digital medium an edition may in fact be a specific view of a larger set of materials.

Scholarly *Edition*

Differentiated from other types of editions, a *scholarly* edition is one that follows scholarly method and purpose, that is undertaken with professional critical judgment and the fullest possible understanding of the relevant primary materials, and that provides clear documentary evidence of the relations and contexts of those primary materials. It is transparent and explicit in demonstrating an attention to the methods of its creation pertinent to the textual situation of its contents and evolving scholarly practice, in documenting the processes by which it was created, and in attending to the concerns of its medium or media. It is typically prepared with an audience of scholars and students in mind, although it may in fact serve a much broader audience, and it may also have pedagogical aims related to how it presents information and supports learning. The rigor of its preparation is ensured through qualitative review, with attention to the application of, or critically constructive relation to, best practices; demonstrated historical knowledge and editorial method; completeness and accuracy of textual account and resultant text or texts; pertinence and utility of textual apparatus and paratext; and other factors relating to its scholarly reliability and usefulness.

A *scholarly* edition is clear about its commitments, and it keeps its promises. It is motivated to support further scholarship through its attention to these principles and their clear exposition, and it is understood to be part of larger scholarly enterprise, ultimately taking its place alongside and possibly in combination with similar works and allowing forms of analysis and engagement beyond those of its editorial intention, supporting further (re)mediation, (re)construction, and (re)mix in the advancement of scholarship in acts that allow, for example, the construction of other editions that may explore alternative hypotheses or challenge notions of authorial intention and editorial authority.

Digital *Edition*

The digital modes in which the scholarly edition of the twenty-first century is so often expressed are deeply significant, but in many cases they serve more to realize potential already inherent in our traditional understanding of the scholarly edition than to overturn that understanding. Although the theme of innovation is common in discussions of digital scholarly editing, it is important to frame that innovation within the context of the goals and overall mission of the editorial enterprise. The digital is neither inherently a site of innovation nor a necessarily useful innovation in itself. In proposing approaches to the assessment and design of effective digital scholarly editions, this statement therefore takes the position that the use of digital methods needs to be carefully thought through, motivated, and explained and that specific digital features need to be consistent with the scholarly goals of the edition (as articulated in the edition's statement of method) instead of serving solely as decoration.

There are specific digital modalities that seem to us to offer particular value for scholarly editions. First, the design of digital editions so that their textual data are captured using standards like TEI opens up important opportunities for alternative deployments of the data: as the basis for other editions, as transcriptions that can be compared using collation tools, as data that can be contributed to a digital repository or aggregated into a text corpus that might support quite different types of analysis. Second, the addressability of digital information (through linked data) makes it possible to support features such as user annotation, commentary, citation, and the creation of additional layers of editorial information. Third, the emphasis on writeability, which is so important to modern digital interfaces, also extends to theories of the digital edition: edition interfaces can serve as environments for manipulation and exploration of the edition's textual space and also as environments within which the user can occupy the role of a contingent editor, examining less-traveled editorial paths and their interpretive consequences. This becomes especially important as one considers the alignment of emerging social computing principles and practices with those traditionally associated with scholarly editing, impacting traditional editorial authority through an emphasis on ongoing open editorial procedure and facilitation.

From another perspective, the digital offers not only additional ways of designing and building scholarly editions but also additional contexts for their use and ways of understanding their pedagogical and cultural importance. Digital communication in general requires changed ideas about literacy, entailing new skills, abilities, and dispositions in front of the activities of reading, writing, and interpreting. Textual scholarship and the study of what N. Katherine Hayles and Jessica Pressman have called "comparative textual media" (vii) lie at the heart of these new literacies, which extend across an ever-expanding variety of textual, visual, and aural media.

Supporting the Scholarly Edition in the Digital World: (Re)Considering CSE's Mandate

Implied in the above, a key pragmatic issue for the CSE, as well as for scholarship more generally in the area, is how we choose to define the terms relating to the past and present of the digital scholarly edition and, indeed, how we choose to view

extant and emerging digital scholarly editions. Many strategies of edition definition follow traditional models and understandings of edition-oriented typology, rooted in work such as that documented and exemplified in Greetham's *Textual Scholarship: An Introduction* (in "Appendix II: Some Types of Scholarly Edition"). These strategies are also evinced in the elements of the CSE's annotated bibliography of key works in the theory of textual editing (Van Hulle). Some definitions focus specifically on the digital, attempting to extend earlier traditions and to typologize digital scholarly editing trends of the past several decades in the context of current and future work (e.g., Siemens et al.; Siemens). Other approaches and examples abound, some listed among materials mentioned in this document and others in and among those scholarly editions in digital form that have been submitted to the CSE for consideration toward the award of the committee's seal that signifies an edition's excellence.

The scope of this statement and indeed the considerations it offers are necessarily framed by the historical moment and the place of the CSE, itself founded with the goal of "improving the state of scholarly editing and . . . encouraging and identifying reliable textual work," a mission that dovetails with larger initiatives within the profession to establish scholarly editing as an authoritative basis for scholarship ("Professional Notes" 274). The current CSE is seeking to further those larger aims at a time when our understanding of terms like *scholarship* and *editing* is under revision. This statement thus serves as an attempt to articulate the CSE's position in relation to that revision process. Pragmatically, we must ask ourselves what we need to know to ensure that the CSE best responds to changes in the field within the scope of its mandate. In brief: What is the frame of address to allow the CSE to best position itself, through the criteria associated with the award of its seal and its potential revision, so as to ensure CSE's continued pertinent function in the scholarly editing community? Digital scholarly editing is an area actively engaged by scholarship at the moment, but reaching relatively stable agreement in the field about it will likely be some years away. There is still much available for us to consider now in relation to the CSE mandate.

In the first instance, to enable us to evaluate editions appropriately we will need to be able to define categories of editions in nonlimiting ways that can be embraced by the CSE and its processes as well, which should reflect the ways in which these considerations are emerging in our community. A key question for us in this regard is, How can we acknowledge the plurality and evolving nature of scholarly editions while nonetheless retaining the ability to recognize excellence and failure, at the same time that we rightly distinguish failure from improper categorization or the limitations of our CSE evaluative model? A clear answer here is that, as per our guidelines, the edition needs to include a statement of purpose that the reviewer can measure against: how appropriate were the methods? how effectively were they carried out?

A further pertinent question is how well our current guidelines, guiding questions for reviewers, and other supporting resources reflect the current and anticipated needs of the future. It has been contemplated that these documents would require revision to accommodate this and other issues. The CSE recognizes that the category of "edition" is extremely broad and sees its own mandate as encouraging and cultivating standards for excellence within that domain. However, at its inception

and in its early documents the CSE adopted a fairly specific definition of the kinds of editions it would cultivate and endorse: critical editions in the tradition of Fredson Bowers. This definition tended to exclude editions of other sorts: for instance, documentary editions. The more recent CSE discussions have emphasized the need to broaden the scope of the CSE's attention to include different editorial modalities. These discussions, however, have highlighted the need for a set of standards of excellence that can generalize well across different types of editions.

To this end, we have arrived tentatively at an initial set of minimal conditions that mark an edition—in our terms—as a scholarly edition now, extant across modalities that could not possibly have been anticipated at CSE's inception or in some cases even a decade ago:

- it must account completely and responsibly for the textual landscape it represents;
- it must fully describe and justify its editorial methods;
- it should reveal the processes by which it was created and disseminated (including data, data structures and constraints, and algorithmic or dynamic processes), and it should include a record of changes and updates made to the edition over time, which otherwise tend to remain invisible in the digital environment;
- it should reveal the judgment and scholarship, the editorial rationales and processes, on which the edition is based;
- it should evince a rigorous standard of accuracy and consistency in applying a particular editorial approach, set of theoretical premises, or method;
- it should demonstrate the appropriate fit among stated methodology, stated goals of the edition (reconstructing authorial intent, reconstructing the social text, etc.), and the nature of the existing textual witnesses;
- it should contain a detailed textual introduction or editorial policy statement, as distinguished from a critical introduction, that outlines these aspects; and
- it should include consideration of how the edition can circulate and function as a scholarly resource over time.

Further conditions that apply specifically to a digital scholarly edition include, but are not limited to, the following:

- it must note its technological choices and be aware of their implications, ideally using technologies appropriate to the goals of the edition (see fit between methods and goals, above), in recognition of the fact that technologies and methods are interrelated in that no technical decisions are innocent of methodological implications and vice versa;
- it should be created and presented in ways ensuring the greatest chance of longevity—addressing this challenge involves infrastructural, financial, and data representation issues (such as the use of widely accepted, open standards);
- it should readily respond to the challenge of maintaining the scholarly ability to be referenced in view of the ways that interfaces change over time; and
- where possible, it should attend to possibilities of sampling, reuse, and remix, supporting approaches to the formation and curation of the edition such as

reconstructing and documenting instances of texts and textual change over time, like algorithmic construction and reconstruction (with possible extensibility, including external data); in doing so, it should attempt to balance considerations for intellectual property and labor with the goals of achieving open access and reusability.

Additional criteria may emerge with further discussion and consideration.

In closing, we wish to reiterate that the CSE remains open to and encourages the practice of a wide variety of editorial approaches, as these relate to both print and digital editions. As its primary mission, the CSE seeks to encourage excellence in scholarly editing, by which we mean above all:

- transparency with respect to data and methods
- clear articulation of motives
- persuasive rationale for the editorial approach taken
- thoroughness and accuracy
- attention to issues of usability of the edition, including questions of audience and of long-term usability

These criteria should be understood as applying equally to print and digital editions. If we take printed books to be "machines of simulation," as Jerome McGann has recently put it (93), this vantage point may help throw into relief the extent to which both types of scholarly edition share overlapping motives, processes, and outcomes, even while we bear in mind their clear and inevitable differences in conception and execution.

For the foreseeable future, at least, there will continue to be editions that exist only or primarily in print as well as those that exist only or primarily in digital form, with the choice of editorial format responding to pragmatic, theoretical, and (in the case of editions affiliated with a university or commercial press) marketing considerations. Print editions will benefit from established practices of marketing and publicity, quality control overseen in part by a press, and proven means of distribution and long-term preservation. At the same time, print editions lack the ability to incorporate seamlessly new discoveries after work has been finished, to make corrections, and to take advantage of the many other features of digital editions discussed in this statement. Print editors would do well to think about the affordances of the digital, including data sharing. Significantly, even print editions now have a digital workflow. But there are also important editorial and interpretive aims that are much more feasible in print than in digital formats, as for instance McGann demonstrates in his discussion of J. C. C. Mays's three-volume *Collected Works of Samuel Taylor Coleridge* (25–30). Building on Hans Walter Gabler's edition of *Ulysses*, Mays's *Coleridge* enables readers to toggle back and forth between "Reading Text" and "Variorum Text," privileging neither through the editorial presentation or apparatus but instead encouraging readers to perceive the poems "in a permanent state of multiple vision" (115). Some materials from print editions might also be made available for nonconsumptive use by scholars in the digital humanities, in ways that would allow

for text analysis and data mining, for instance. High-quality editions thus continue to be produced in both print and digital forms. The CSE will remain committed to encouraging and discerning the best practices for multiple types of editors and with multiple categories of readers in mind.

Works Cited

Greetham, D. C. *Textual Scholarship: An Introduction*. New York: Garland, 1994. Print.

Hayles, N. Katherine, and Jessica Pressman. "Introduction: Making, Critique: A Media Framework." *Comparative Textual Media*. Ed. Hayles and Pressman. Minneapolis: U of Minnesota P, 2013. vii–xxxiii. Print.

Mays, J. C. C., ed. *Collected Works of Samuel Taylor Coleridge*. 3 vols. Princeton: Princeton UP, 2001. Print.

McGann, Jerome. *A New Republic of Letters: Memory and Scholarship in the Age of Digital Reproduction*. Cambridge: Harvard UP, 2014. Print.

"Professional Notes and Comment." *PMLA* 95.2 (1980): 264–76. Print.

Siemens, Ray. "Disparate Structures, Electronic and Otherwise: Conceptions of Textual Organisation in the Electronic Medium, with Reference to Electronic Editions of Shakespeare and the Internet." *Early Modern Literary Studies* N.p., 1998. Web. 13 Aug. 2015. <http://purl.oclc.org/emls/03-3/siemshak.html>.

Siemens, Ray, et al. "Toward Modeling the Social Edition: An Approach to Understanding the Electronic Scholarly Edition in the Context of New and Emerging Social Media." *Literary and Linguistic Computing* 27.4 (2012): 445–61. Web. 12 July 2012. <http://dx.doi.org/10.1093/llc/fqs013>.

Van Hulle, Dirk. *Electronic Textual Editing: Annotated Bibliography: Key Works in the Theory of Textual Editing. Text Encoding Initiative*. TEI, 31 Oct. 2007. Web. 13 Aug. 2015.

This statement was issued by the Committee on Scholarly Editions in September 2015 and approved by the MLA Executive Council at its October 2015 meeting.

7-1-2009

Edition, Project, Database, Archive, Thematic Research Collection: What's in a Name?

Kenneth M. Price
University of Nebraska - Lincoln, kprice2@unl.edu

DHQ: Digital Humanities Quarterly

Preview
Summer 2009
Volume 3 Number 3

Edition, Project, Database, Archive, Thematic Research Collection: What's in a Name?

Kenneth M. Price <kprice_at_unlnotes_dot_unl_dot_edu>, University of Nebraska-Lincoln

Abstract

What are the implications of the terms we use to describe large-scale text-based electronic scholarship, especially undertakings that share some of the ambitions and methods of the traditional multi-volume scholarly edition? And how do the conceptions inherent in these choices of language frame and perhaps limit what we attempt? How do terms such as edition, project, database, archive, and thematic research collection relate to the past, present, and future of textual studies? Kenneth M. Price considers how current terms describing digital scholarship both clarify and obscure our collective enterprise. Price argues that the terms we use have more than expressive importance. The shorthand we invoke when explaining our work to others shapes how we conceive of and also how we position digital scholarship.

What are the implications of the terms we use to describe large-scale text-based electronic scholarship, especially undertakings that share some of the ambitions and methods of the traditional multi-volume scholarly edition? What genre or genres are we now working in? And how do the conceptions inherent in these choices of language frame and perhaps limit what we attempt? How do terms such as *edition*, *project*, *database*, *archive*, and *thematic research collection* relate to the past, present, and future of textual studies? Drawing on a range of resources, including the *Walt Whitman Archive*, I consider how current terms describing digital scholarship both clarify and obscure our collective enterprise. In addition, I will use the final term, *thematic research collection*, to discuss yet-to-be-developed parts of the *Whitman Archive* dealing with place-based cultural analysis and translation studies as a way to illustrate the expansive possibilities of this new model of scholarship. 1

Digital textual studies seem to me inadequately described by the terms now available. *Project* is amorphous; *archive* and *edition* are heavy with associations carried over from print culture; *database* is both too limiting and too misleading in its connotations; and *digital thematic research collection* lacks a memorable ring and pithiness. The terms we use have more than expressive importance. The shorthand we invoke when explaining our work to others shapes how we conceive of and also how we position digital scholarship. We need a new term that is vivid enough to be memorable, elastic enough to cover a class of like things, and yet restrictive enough to allow us to include some scholarly undertakings and not others. Ordinary readers and academics alike rely heavily on the work of editors, yet the standing of editors in the academy has for decades been shaky at best. For many people, electronic work is even more dubious: what relatively short history it has is marked by distrust, denigration, and dismissal. We all know the charges, however distorted they may be: digital work is ephemeral, unvetted, chaotic, and unreliable. When suspicion of the value of editing combines with suspicion of the new medium, we have a hazardous mix brewing. There is a danger that if humanities scholars do not undertake the key work of textual transmission, this work will be done by librarians and systems engineers — that is, it will be done by people with less specialized knowledge of the content. In the fraught circumstances of the academy, driven by a prestige economy, humanities scholars are well advised to be highly self-conscious about what we do and how we describe it. 2

Edition

What do we mean when we use the term *edition*? Even among print editions, there are a number of variations: selected editions, reader's editions, and some boldly claiming to be authoritative or definitive editions. The descriptive word "scholarly" has been applied to numerous approaches: authorial or social, critical or documentary, genetic, eclectic, or best text [Stauffer 2007]. Successful scholarly editions yield a text established on explicitly stated principles by a person or a group with specialized knowledge about textual scholarship and the writer or writers involved. What makes the edition scholarly, of course, is the rigor with which the text is reproduced or altered and the expertise deployed in the offering of suitable introductions, notes, and textual apparatus. 3

For those of us who work on prominent figures who have received previous treatment, our own textual work intervenes in an ongoing editorial tradition. A fundamental and often vexingly difficult question is, what should go in an edition? Like most digital editing endeavors, the *Walt Whitman Archive* must proceed with an awareness of the print past — in our case, especially of two significant attempts to present Whitman in scholarly editions: *The Complete Writings of Walt Whitman* (G. P. Putnam's Sons, 1902) and *The Collected Writings of Walt Whitman* (New York University Press, Peter Lang, and the University of Iowa Press, 1961-2004). This awareness produces competing impulses: we want to benefit from and respond to past work, but we also want to avoid constraints on thought and action that were a result of print-based limitations. As editors, we acknowledge the ways of knowing that are enabled by our predecessors — they are the cultural history we inherit — but our job is also to extend their efforts and to produce new ways of knowing that are responsive to cultural, critical, and technological changes (as well as the discovery of documents and the development of new biographical insights) that have happened in the interim. 4

The language of the *Walt Whitman Archive*'s first grant application to the National Endowment for the Humanities (NEH), drafted in 1999, shows how we were thinking of our digital work as in dialogue with the print past. We wrote, 5

> Our goal has been to build upon the strengths of the *Collected Writings* edition, most volumes of which were supported by grants from the National Endowment for the Humanities. The amount of Whitman's work is so huge that no two scholars could hope to edit it effectively in a lifetime — fourteen scholars spent the better parts of their careers editing the materials that now make up the *Collected Writings*. But we do believe that developments in electronic scholarship have made it possible to enhance and supplement the *Collected Writings* by editing the materials that have not yet been included (and adding the materials that have come to light since the *Collected Writings* volumes were issued) and by digitizing and encoding the *Collected Writings* so that these disparate volumes — which often arrange material in confusing and contradictory ways — can function seamlessly and so that Whitman's materials can be presented effectively in any number of new configurations: by genre, by date, by keyword, by subject. The electronic environment can also allow us to make available not just printed transcriptions of Whitman's manuscripts, letters, and books, but to deliver actual facsimile images of the original documents. [Folsom and Price 1999]

It would be fair to acknowledge that Ed Folsom, co-director of the *Whitman Archive*, and I have had evolving views of the relationship between our undertaking and its most recent print predecessor, the *Collected Writings of Walt Whitman*. Our gradually shifting views have been shaped in part by discussions with publishers. At various times, we considered entering into agreements with two publishers — Primary Source Media and the University of Virginia Press — and in fact reached late stages of contract negotiations with each of them. Initially, we reasoned that if a publisher could secure the permissions for us to use the copyrighted material in the twenty-two volumes of the *Collected Writings* published by New York University Press, a significant amount of work, some of it meticulously done, could be preserved and extended.[1] Of course this line of thinking raised a key issue: if a new publisher had to pay for the permissions, the site, or some significant part of it, would need to be commercial in order to recover these and other costs, and perhaps make a profit as well. We were not absolute purists committed always to building a completely free site. In fact, there were extended periods when we were convinced that such an approach would not be possible for a poet like Whitman who left so much debris everywhere. We thought that editing such chaos would demand the combined resources and know-how of the scholarly, library and archival, and publishing communities. Gay Wilson Allen, a general editor of the *Collected Writings*, commented about editing Whitman, "Sometimes his exhausted editors almost wish that he had had two or three good house fires, and considering the houses he lived in, it is also astonishing that he did not" [Allen 1963, 8]. 6

For us, then, a key question emerged: would we conceive of the *Whitman Archive* primarily as being the remediation of the *Collected Writings*? We recognized that our relationship to the *Collected Writings* is problematic: we have a half-century of valuable editorial work collected there, but the limits of a print format make this edition a trial to use. The *Collected Writings* has been the standard edition, the edition cited by American literary scholarship over the past few decades, but much of the work needs to be done again and the presentation re-conceptualized. We struggled to come to terms with a giant from the print past. And yet this monumental edition was both enormous and characterized by some inexplicable omissions, most notably Whitman's revelatory poetry manuscripts. As our initial grant application pointed out, 7

> [W]e have Whitman's laundry lists in print; we have the business cards of his sidewalk repairman in print, but we don't have the manuscripts of "Song of Myself" in print... His poetry manuscripts and periodical publications reveal, among other surprising things, a Whitman who devoted extraordinary time and care to the creation of a poetry that appeared to be quick and spontaneous; his manuscripts expose an artist whose casual, loafing persona was in fact the result of intensive and obsessive artistic labor. [Folsom and Price 1999]

In retrospect, it is clear that we have responded to the *Collected Writings* not by "digitizing and encoding" it but by prioritizing work on material not included there: photographs, bibliography, full texts of various editions of *Leaves of Grass*, archival guides to manuscripts, transcriptions of manuscripts, contemporary reviews of Whitman's writings, and 8

so on. If we were the first editors of Whitman, this order of development for an online resource would have been peculiar. Certainly some of Whitman's prose, *Democratic Vistas* or *Specimen Days*, for example, or his correspondence might rank ahead of some of these items in most people's sequencing list. But of course we do work within a historical context, and what has seemed most pressing (and perhaps most fundable) have been those things altogether neglected or poorly treated by the *Collected Writings*.

9 Sometimes we learn to be thankful for our failures, and I am certainly grateful now that our negotiations with publishers always went bust. I think — because of a recent NEH challenge grant to be discussed later — that the *Whitman Archive* is in an unusual position: we now have a team of people and the resources in place so that, with reasonable luck, we ought to be able to achieve a more expansive *Whitman Archive* than the already quite extensive site, and to keep it freely available. There are of course examples of other large, not to say gargantuan, free sites. But we should not underestimate the challenges attendant on making vast amounts of material freely available since "free" means no cost to the end user, not the creators.

10 It is reasonable to wonder why Whitman needs to be edited if there have been two previous scholarly editions. And it is reasonable to acknowledge, in response, motivations that have nothing to do with the electronic medium specifically. Editorial work is one way to engage in historical criticism and to help bring the past into the present so it may live in the future. Although the shelf life of a scholarly edition far exceeds that of a monograph, scholarly editions begun half a century ago for Whitman in one case, or a century ago in the other, now seem inadequate. Their approaches require rethinking, not to mention the need to add material and convey new discoveries. Editions of modern writers are almost always selective. Still, a selection ought to include the most important items. If asked to pick Whitman's most important single text, many would name the first publication of *Leaves of Grass* (1855). Here Whitman was at his boldest and most experimental, and the book has elicited some memorable reactions over the past 150-plus years: Ralph Waldo Emerson found it to be "the most extraordinary piece of wit and wisdom that America has yet contributed" [Emerson 1938-1994, 446]. William Carlos Williams called the first *Leaves* "a book as important as we are likely to see in the next thousand years" (Williams, quoted in Hindus 1955, 3) . Clearly, this is a highly significant book. And we might expect the 1855 *Leaves* to be the highlight of an edition of Whitman's writings. Strangely enough, neither *The Collected Writings of Walt Whitman* nor the earlier *Complete Writings of Walt Whitman* bothered to include it.

11 How do we explain this omission? To a large extent, this odd result stems from twentieth-century editorial practices for establishing authoritative or definitive texts that encouraged the selection of a single text. The economics of print publishing — combined with the dominant editorial theories of the mid twentieth-century — made the so-called deathbed edition of *Leaves of Grass* the one most commonly featured in various commercial and scholarly editions. That final authorized printing of Whitman's book is in fact presented twice in the New York University Press edition: it serves as the basis of both the Comprehensive Reader's Edition and the *Leaves of Grass* Variorum. The deathbed edition is remarkable, but it could not be described as Whitman's most daring, most experimental, or even most coherent volume.

12 Print editions of Whitman tended to falter when dealing with multiplicity, whether of versions or of authorship. Whitman is well known as the writer who couldn't stop writing, revising, and reissuing *Leaves of Grass* (a book that appeared in six radically distinct American editions in his lifetime). Less well known is Whitman's involvement in collaborative enterprises. In fact, when we think of the great collaborators in literary history, Whitman hardly jumps to mind. Instead, we remember that Whitman was so self-reliant that for the first edition he more or less did everything: wrote the poetry, designed the book, set some of the type, distributed the book, and anonymously reviewed it. He appears to be dead set against even the largely invisible and ordinarily neglected forms of social authorship, a poet acting out the role of the solitary singer made famous in "Out of the Cradle Endlessly Rocking." Yet this poem also dramatizes collaboration, with one set of voices prompting another, bird song and human song, a single trill and the thousand responsive chords from a thousand different singers to follow.

13 Arguably the medium of print itself encouraged earlier editors to take a restricted view that often remained blind to the social aspects of textual production. It is easier, frankly, to exclude contributions made by book designers, copyeditors, typesetters, and others. Yet if we think longer and harder about Whitman's own career, the extent of his collaboration — almost entirely ignored by *The Collected Writings of Walt Whitman* — is striking. Whitman collaborated with typesetters, designers, and proofreaders, as he readily acknowledged,[2] and also in his journalism, both as editor and writer; in his extensive though anonymous contributions to the early Whitman biographies by R. M. Bucke and John Burroughs; in heretofore uncollected interviews (now being edited by Brett Barney); in his extensive conversations with Horace Traubel — a 5,000 page trove of information. In fact, his correspondence itself is fundamentally a collaborative undertaking involving (ordinarily) two-way engagements, though the strong authorial bias of the *Collected Writings* is clear in their featuring of just Whitman's outgoing correspondence.

Project

14 *Project* is a bigger, baggier term than edition and is far less specific in what it suggests about the type of work being undertaken. *Project* can describe everything from fixing a broken window on the back of a house to the Human Genome Project. In a literary context, *editions* and other results tend to emerge out of *projects*, but what constitutes the *project* is also the entirety of the undertaking: space, personnel, atmosphere, and the totality of all efforts. An *edition* might result

from a project, without being the *project*, which includes all of the work conducted and records produced. The *Whitman Archive*, when regarded as a project, encompasses the compiled email discussion list that fitfully records the building of the *Archive* and the thinking that has gone into it. The documentation of a project, in our case, includes the behind-the-scenes Works-in-Progress page, with its assortment of information, including grant proposals, minutes from Whitman planning meetings over the years, a manuscript tracking database, an image warehouse, and project-related humor.

15 *Project* is not a favored word in every context. When I sent drafts of a "We the People" challenge grant application to NEH program officers, I was struck by how forcefully they discouraged me from the using the word *project*, at least in the context of that competition. Their reasoning was that challenge grants were intended to fund permanent entities, unlike a *project* which they conceived of as having a finite temporal life. For me, "Whitman Project" and "*Whitman Archive*" were more or less interchangeable terms. I had to make a real effort to purge the document of all references to *project*. It was a neutral term to me: *project* was so natural as to be almost invisible in the drafts and certainly did not raise a red flag.

16 This story raises a larger issue: what happens when an undertaking becomes not just rhetorically but practically open-ended, when it has the good fortune or obligation to be an ongoing concern? We were successful with our challenge grant application, and we are now well along in building a $2 million permanent endowment for the *Whitman Archive*. Thanks to this remarkable turn of events, the *Whitman Archive* can now plan on an ongoing annual budget comparable to what one might expect annually from a major two- or three-year grant from a federal agency or foundation. And, remarkably, in this case, there is no end date to that support.

17 For the 2007 Digital Humanities conference at the University of Illinois, Urbana-Champaign, Matt Kirschenbaum coordinated a panel called "Done. Finished Projects in the Digital Humanities." He asked, "How do we decide when we're done? What does it mean to finish something? How does the 'open ended nature of the medium' (a phrase we all pay lip service to) jibe with the reality of funding, deadlines, and deliverables? What can we learn from finished projects, both successful and unsuccessful? For that matter, how do we define success and failure? Are 'we' the ones who ought to be defining it? If not, who?" These are good questions, and at the *Whitman Archive* we find ourselves concerned with them even as we face different considerations as well. What happens when work plans realistically *could* continue over generations? What is the best way to plan for that type of future?[3] A theoretical possibility of digital scholarship — the indefinite expansibility — has become a lived reality in our case. We are only now absorbing the meaning of this grant, but one implication is that it provides us with the license, perhaps even the charge, to be as bold and ambitious as our talents and energies allow.

Database

18 How adequate is the term *database* for describing the type of large scale electronic projects we have been considering? Throughout this essay, I have used the *Walt Whitman Archive* as a testing point and illustrative example. To discuss the *Whitman Archive* in terms of database is especially timely now because *PMLA* recently featured an article about the *Walt Whitman Archive* by Ed Folsom, "Database as Genre: The Epic Transformation of Archives," and included a handful of responses (along with Folsom's response to the responses). The ensuing discussion made clear that people understand the term *database* in a variety of ways and attach different connotations to the word. These differences arise mainly from a distinction between 1) a strict definition of *database* — as a technical term in an electronic context database refers primarily to a collection of structured data that is managed by a database management system, most commonly based on a relational model; and 2) a looser use of *database* that employs the term on a more metaphorical level.

19 As the *PMLA* discussion of the *Whitman Archive* indicates, *database* can be a suggestive metaphor because it points to the re-configurable quality of our material (and that of similar sites). The term also conveys simultaneously "finished" and "unfinished" qualities; while a *project* can be logically thought of as "done" or "not yet done," we usually conceive of a database as usable as soon as it begins to exist, and we take as a given that the data will continue to proliferate, potentially indefinitely. The *Whitman Archive* resembles a database in that its content is discrete computer files that function atomistically: as functional units within a computing system each item is just as important as every other item.

20 If the *Walt Whitman Archive* resembles a database (without meeting the specifications of a technical or a literal definition), so, too, does Whitman's own process of composition. As Folsom notes, "Whitman formed entire lines as they would eventually appear in print, but then he treated each line like a separate data entry, a unit available to him for endless reordering, as if his lines of poetry were portable and interchangeable, could be shuffled and almost randomly scattered to create different but remarkably similar poems" [Folsom 2007, 1574-75]. At times, it almost seems as if Whitman were anticipating Raymond Queneau's *Cent Mille Milliards de Poèmes* [One Hundred Thousand Billion Poems], a fascinating book in which the pages are cut horizontally so that each verse in each sonnet of the collection can be turned separately and all combinations of choices are poetically grammatical. (Queneau estimated that a reader would have to spend two hundred million years, working twenty-four hours a day, to read every combination.[4]) Whitman's own cutting and pasting of lines, and his rearranging of poems to make other poems is not this extreme — nor is it as extreme as Samuel Beckett's experiments in *Lessness*[5] — though there is some resemblance to both. Finally, though, what may appear random ordering in Whitman is best understood as restless experimentation, a

combinatory and recombinatory poetics, guided by Whitman's recurrent drive to improve the effectiveness of his poems. Here, for those willing to use the term *database* metaphorically and to recognize non-electronic forms of databases, we can think of database as a key tool for Whitman himself: his storehouse of poetic lines, in both manuscript and print, was his working database for future compositions, one that he had always only partial access to because of the scattering of his documents but that nonetheless served as a means of composition.

21 If we turn to more literal uses of the word database and think about the *Whitman Archive*, we see that it is a complex composite structure that includes numerous databases and XML files. Folsom's description of the *Whitman Archive* as "a huge database" is illuminating when taken metaphorically, though it is less helpful when taken literally, because the entirety of the *Whitman Archive* is not a single database any more than it is, as Jerome McGann asserts, merely XML files plus XSLT. In fact, the *Walt Whitman Archive* is comprised of numerous databases (some public and some not) along with many XML files including TEI, EAD, and XHTML files.[6] McGann goes on to claim that the XML and XSLT work together to "allow users to access and—through an X-query-based search engine—manipulate *The Walt Whitman Archive* in the ways that Folsom rightly celebrates" [McGann 2007, 1588]. Ironically, though, in the course of denying the applicability of *database* as a term suitable to the *Whitman Archive*, McGann overlooks that our search engine is entirely dependent on translating the XML files into database form. At a more general level, McGann is perceptive in noting that any database represents an initial interpretation of the material. A database is not an undifferentiated sea of information out of which structure emerges. Argument is always there from the beginning in how those constructing a database choose to categorize information — the initial understanding of the materials governs how more fine-grained views will appear because of the way the objects of attention are shaped by divisions and subdivisions within the database. The process of database creation is not neutral, nor should it be.

Archives and Digital Thematic Research Collections

22 Having discussed *edition*, *project*, and *database* separately, I now turn to consider the final two terms together, *archive* and *digital thematic research collection*. In the past, an archive has referred to a collection of material objects rather than digital surrogates. This type of archive may be described in finding aids but its materials are rarely edited and annotated as a whole. In a digital environment, *archive* has gradually come to mean a purposeful collection of surrogates. As we know, meanings change over time, and *archive* in a digital context has come to suggest something that blends features of editing and archiving. To meld features of both — to have the care of treatment and annotation of an edition and the inclusiveness of an archive — is one of the tendencies of recent work in electronic editing. One such project, the *William Blake Archive*, was awarded a prize from the Modern Language Association recently as a distinguished scholarly edition.[7]

23 Digital archives are often notable for their depth and breadth of coverage of whatever the stated thematic interest is. Such scope has not been common in editing. Indeed it is possible to see a tension in the very term *collected edition* because collecting and winnowing are two very different activities. Thomas Wentworth Higginson, in a review of the *Complete Writings of Walt Whitman*, might have been commenting on the *Whitman Archive* when he wrote, "[T]he present editors do not shrink from inserting not only the details of every change, but even the unprinted variations which have hitherto existed in manuscript only" [Higginson 1903, 400]. Of course, the more inclusive an edition becomes the more it may be dominated by the surviving "discarded" writings, especially for writers who kept many documents [Folsom 1982, 374]. Some feel that we do violence to the wishes of writers when we make their second-rate material available to the public, while others celebrate what they believe is made possible by inclusive editions: a new, deepened, and enriched sense of the artist's process of composition, preoccupations, and achievements. Ultimately, the whole question of what is in keeping with the wishes of a writer is beside the point. We do not edit for writers themselves but for our own purposes as scholars and readers.

24 Peter Shillingsburg expresses skepticism about the advantage of the archival approach:

> The computer makes possible, we are told, the juxtaposition of all the relevant texts in their linguistic and bibliographic variant forms. Thus a library of electronic texts, linked to explanations and parallels and histories, becomes accessible to a richly endowed posterity. To the extent that such archives contain accurate transcriptions, high resolution reproductions, precise and reliable guides to the provenance and significance of their contents, and the extent to which they are comprehensive, to that extent they are "definitive" — until the next generation of critics and scholars with new interests notices some other aspect of texts that scholarly editors of the past (by then that will be us) took for granted and ignored. But already, information overload has set in. The comprehensiveness of the electronic archive threatens to create a salt, estranging sea of information, separating the archive user from insights into the critical significance of textual histories. [Shillingsburg 2006, 165]

Shillingsburg focuses on the limits of a form still being developed as opposed to the potential of that form. Nothing in the archive form intrinsically requires it to be "estranging" or alienating, of course. An electronic archive can be as welcoming as fresh water and as rewarding as the wit of its creators can make it. Having a lot of information is not

inherently more estranging than having less information. Nothing guarantees the effectiveness of selective treatment accompanied by "textual histories," and nothing guarantees effectiveness of more comprehensive treatment accompanied by textual histories. In each case, everything depends on the quality of the editorial work. Digital and print scholarship are equally embedded in history, and both share a vulnerability to aging.

25 Another term that is more or less synonymous with electronic *archive* is *digital thematic research collection*.[8] Some prefer this term because it may avoid some of the misleading connotations of *archive* — ordinarily people assume that materials in a traditional print-based archive are unedited.[9] Carole Palmer writes about *thematic research collections*,

> Collections of all kinds can be open-ended, in that they have the potential to grow and change depending on commitment of resources from collectors. Most thematic collections are not static. Scholars add to and improve the content, and work on any given collection could continue over generations. Moreover, individual items in a collection can also evolve because of the inherent flexibility (and vulnerability) of "born digital" and transcribed documents. The dynamic nature of collections raises critical questions about how they will be maintained and preserved as they evolve over time. [Palmer 2004, 351]

Archive is a self-designated term, one adopted by the creators of resources. In contrast, *digital thematic research collection* is a term used by people describing the work created.

26 *Thematic research collection* may be the most accurate term for what many of us are attempting, but it has not gained currency because it is neither pithy nor memorable. Carole L. Palmer notes that a *digital thematic research collection* is the closest thing to the laboratory that we have in the humanities — the place where necessary research materials are amassed. I have argued elsewhere that in a "digital context, the 'edition' is only a piece of the 'archive', and, in contrast to print, 'editions', 'resources', and 'tools' can be interdependent rather than independent" [Price 2007, 435].

27 Does collecting — the emphasis in Palmer's description — qualify as research, as a scholarly genre? A digital thematic research collection possesses the virtues of a traditional scholarly edition while containing much more. We may nonetheless wonder about how helpful the term *digital thematic research collection* is to the uninitiated. Nothing in the term indicates editorial rigor and nothing points to the value added by scholarly introductions, annotations, and textual histories. The only thing that seems to separate it from a mass digitization project is the "thematic" element. However, one can imagine a mass digitization project that is thematic and that lacks editorial supervision and intervention in the reader's experience of the text. Can we find a better term that indicates this difference? Does *digital thematic research collection* communicate its meaning adequately?

28 If literary scholars who are assembling electronic texts are becoming fundamentally or solely "literary-encoders" and "literary-librarians," then, despite my own recognition of the inseparability of interpretation and encoding, I fear for the standing of their work when judged by faculty in humanities departments (Schreibman, as quoted in [Palmer 2004, 352]). Without care and forceful practical examples and theoretical essays, the same prejudices and misunderstanding that drove editing and bibliography from the center to the periphery of literary studies will continue to prevail. We also need descriptions of digital thematic research collections that highlight the editorial work and other types of scholarly value that are added to the raw materials populating the collection. In many circles, editing — whether it is print-based or electronic — is regarded as pre-critical work. Some editorially related tasks are fairly routine and do not require scholarly expertise (the same is true of critical work as well). And yet others clearly do, and we need to find ways to clarify how historical knowledge, theoretical sophistication, and analytical strengths are necessary to the creation of a sound text or texts and accompanying scholarly apparatus in a successful edition.

29 Some components of a digital thematic research collection or archive may stretch ordinary understandings of *edition*. Many thematic research collections or archives aim toward the ideal of being all-inclusive resources for the study of given topics. A good thematic research collection might begin with an edition conceived in inclusive terms. Digital thematic research collections go far beyond traditional editions in their presentation of many types of materials. They are often even more "organic" than print editions (despite their technological aspects) — that is, they grow, evolve over time, based very much on immediate circumstances. For the *Walt Whitman Archive*, new work on the Civil War is now underway because an expert on Abraham Lincoln at the University of Nebraska-Lincoln, Kenneth J. Winkle, and I perceived a scholarly need and are interested in collaborating on this undertaking. New work on translation — I will say more about both of these new endeavors later — developed because Matt Cohen, already associated with the *Whitman Archive*, was interested. Being published online but being simultaneously a work-in-progress allows for a flexibility in the *Whitman Archive* that print editions could never have. New scholars with new ideas may emerge at any time, creating new and unexpected additions to our work.

30 I mentioned earlier that the theoretical possibilities of digital scholarship might oblige us to boldness — the present moment, when electronic scholarship is still nascent and the boundaries are still capable of being moved, provides a mandate to innovate and expand possibilities. Ideally, a digital thematic research collection would also allow for the study of cultural contexts. In the case of Whitman, we might want to study him as a city poet. He once said that *Leaves of Grass* "arose out of my life in Brooklyn and New York from 1838 to 1853, absorbing a million people, for fifteen years

with an intimacy, an eagerness, an abandon, probably never equaled" (quoted in Reynolds 1995, 83) . A life-long city-dweller, his work also emerged out of New Orleans, Washington DC, and Philadelphia/Camden, New Jersey. We would like for the site to enable and to promote interpretations of place-based writing that were not possible before. It would be useful to be able to study all of these areas with dynamic maps containing detail down to the block level. Period maps exist for Washington, DC, New York, Brooklyn, Philadelphia, and New Orleans. New discoveries will emerge once we can ask different questions because of having a great deal more information from census records, maps, health records, police reports, possibly even information on sexual subcultures, and so on.

I have recently begun work on a digital undertaking that may or may not become part of the *Whitman Archive*. Whether the project ultimately is folded into the *Archive* or remains a separate, stand-alone collection, it certainly grew out of my work on the *Archive*. We might think about it as budding off of an existing digital thematic research collection and taking on a life of its own. The project "Civil War Washington: Studies in Transformation" draws on the methods of many fields — literary studies, history, geography, computer-aided mapping — to create an experimental digital resource. The President and the poet both experienced the War from vantage points in the nation's capital, Lincoln striving to reunite the divided nation and Whitman caring for tens of thousands of wounded soldiers. Their activities and perspectives chronicle the War and provide insights into the large and complex forces that transformed Washington from a sleepy Southern town to the symbolic center of the Union and nation. 31

We are gathering uncollected factual data about an urban space that served as the center both of the Union's War effort and of a divided nation, where hospitals arose overnight, wounded men moved in and out, "contraband camps" of fugitive slaves developed, and temporary shelters were erected to house the city's swelling population, which tripled during the four years of the War. Washington was a noisy city during these years: the noise in the city was of construction as work on the Capitol continued; the noise just outside the city was of destruction as the Confederate army worked to tear it down. Even as bridges were defended and a ring of forts made this space the most heavily defended city on earth, Washington fostered vibrant life. 32

"Civil War Washington: Studies in Transformation" will situate Lincoln and Whitman in the midst of a rich field of geo-spatial and temporal data. At the heart of the project will be richly layered, interactive maps plotting both geographic and temporal data that clarify the transformation of Washington, DC. The maps and underlying databases will make it possible to analyze change over time as structures grew and the population swelled and developed a new ethnic and racial mix. We will make possible multifaceted and dynamic studies of Lincoln's and Whitman's activities during the War years, based on textual and statistical evidence and using the power of maps and graphs to illustrate historical change. Lincoln's and Whitman's routes can be plotted on a daily and sometimes hourly basis. We believe that by providing a rich backdrop of census, health, and hospital records; theater schedules; horsecar routes; and other factual data, we will make possible a better understanding of Lincoln's and Whitman's lives and their roles in the transformation of the nation and its capital. 33

Another extension of the *Whitman Archive* now being undertaken serves to expand trans-linguistic, cross-cultural understandings. Whitman scholarship offers rich opportunities because *Leaves of Grass* has been translated into every major language. One of the *Archive*'s objectives is to present editions of Whitman's work key to literary, cultural, and historical study of the poet and his work's effects. Thus Matt Cohen has taken the lead in tackling a digital edition of the first extensive translation of his work into Spanish. Álvaro Armando Vasseur's 1912 selection from *Leaves of Grass* is the work of a Uruguyan poet who translated Whitman not directly from English but via an earlier Italian translation. This fascinating text tells us a lot about the circulation of culture. Making a version of *Leaves* available to the Hispanaphone world seems fitting given current trends in U.S. demographics and in light of the many calls to internationalize American studies. 34

We supplement the translation with a critical introduction and a sample back-translation into English in order to give those unable to read Spanish an opportunity to see how the text was altered in the process of translation. For example, consider the following lines as given by Whitman: 35

> The disdain and calmness of martyrs,
> The mother of old, condemn'd for a witch, burnt with dry wood, her children gazing on,
> The hounded slave that flags in the race, leans by the fence, blowing, cover'd with sweat,

And here is how Vasseur rendered these lines as revealed in a literal back translation: 36

> The mother of old condemned as a witch and burned over dry firewood, before her children's eyes,
> The slave, persecuted like an imprisoned woman, who falls mid-flight, all atremble and sweating blood.

Vasseur's direct comparison of the slave to a woman presumably is based on their common lack of power, but it also creates some cross-gendered possibilities that turn the passage in new ways. Whitman had distinct units — separate lines — for the witch and the hounded slave. An association could be made between them because of their juxtaposition, but that association is hardly insisted on in the English original. Vasseur turns the suggestion of a link into an unmistakable link. Now racial slavery has become associated with the irrationality of the inquisition and serves to remind the reader of the widespread support of slavery by the church in the U.S. (and in South America). While this 37

reading is only barely available in Whitman's original, in Vasseur's translation it appears on the surface. This passage clarifies that translating a text is interpreting it in another language. To ignore such interpretations is to ignore an enormous part of Whitman's reception in the world.

38 We have either in progress or the planning stages work on Whitman and other languages (German, Russian, Ukrainian, Portuguese, and Chinese). This will begin to better place him in a world context rather than situtating him solely in Anglophone culture. The work will provide valuable texts to further Whitman studies and through associated commentary reflect the social, historical, and linguistic milieus of the nations in which the translations were done, thereby once again stretching the bounds of what a digital thematic research collection originally envisioned within much narrower parameters can do. These possibilities, the ever-emerging questions and new directions, go far beyond the ordinary edition in the pre-digital age.

39 As I have indicated, we do not have an adequate term to describe the digital scholarly work now underway in numerous projects. What is it that we want our descriptive word to capture: is it the physical thing? Digital sites, contrary to popular (and sometimes scholarly) opinion, are physical things after all — they take up space, can be created and destroyed, and so on. Is it the nature of the content? If so, we need a word that suggests what can be an infinitely extensible resource. Or should we emphasize, primarily, the way we make the thing, the collective that has come together in order to do work on a new scale in humanistic study?

40 Importantly, we should not strive to fit our work to one or another existing term but instead expect that, in time, terms will alter in meaning — or new ones will come into existence — so as to convey the characteristics of a new type of scholarship. I strongly agree with Peter Shillingsburg that a new term is needed, though I am not enthusiastic about his proposed term: *knowledge site*. (So many places and institutions could justifiably be called *knowledge sites* that the term seems unlikely to become identified with a particular genre of electronic scholarship.) I propose instead a not-immediately-intuitive but perhaps ultimately more promising alternative: *arsenal*.[10] The online etymological dictionary helps explain the appeal of the term:

> *arsenal*
>
> 1506, "dockyard," from It. arzenale, from Ar. dar as-sina'ah "house of manufacture, workshop," from sina'ah "art, craft, skill," from sana'a "he made." Applied by the Venetians to a large wharf in their city, which was the earliest meaning in Eng. Sense of public place for making or storing weapons and ammunition is from 1579.

I like the emphasis on workshop since these projects are so often simultaneously products and in process. I also like the stress on craft and skill, a reminder that editing is not copyist work. The "public place for making" suits current aspects of the genre under discussion and will no doubt characterize it even more in this age of social networking. The dockyard connotations of *arsenal* are helpful in suggesting a kind of inclusiveness about all the vessels, sloops, ketches, and yawls that can hook up to it. (The wharf and dockyard are places of multilingual exchange.) The obvious objection to the term *arsenal* is that it seems militaristic in current usage. Yet we should recall that *magazine* once primarily meant a storehouse for weapons and ammunition. If the primary meaning of *magazine* can shift from being a storehouse of weapons to a storehouse of mixed content for periodical publication, who knows what could happen with *arsenal*?[11] We are, for better or worse, always entangled with force and power: the Internet itself has its origins in the military. Perhaps one step toward turning swords into plowshares is to seize a word like *arsenal* and make it our own. Can we imagine a world in which what is emphasized is not the created thing so much as the group of people who are now joined together for a common purpose?

Notes

[1]After New York University Press published twenty-two volumes of *The Collected Writings of Walt Whitman*, the publishing house of Peter Lang published two additional volumes of Whitman's journalism, and the University of Iowa Press published a single supplemental volume of Whitman's correspondence.

[2]After the publication of the 1881-1882 *Leaves of Grass*, Whitman remarked, "All this is not only my obligation to Henry Clark, but in some sort to all proof-readers everywhere, as sort of a tribute to a class of men, seldom mentioned, but to whom all the hundreds of writers, and all the millions of readers, are unspeakably indebted. More than one literary reputation, if not made is certainly saved by no less a person than a good proof-reader. The public that sees these neat and consecutive, fair-printed books on the centre-tables, little knows the mass of chaos, bad spelling and grammar, frightful (corrected) excesses or balks, and frequent masses of illegibility and tautology of which they have been extricated" [Whitman 1978, 256]

[3]Issues involving long-term preservation come to mind, of course. A simple curation is not viable. That is, we cannot hand over to a library the present-day *Whitman Archive* and expect people fifty years from now to find its interface and technical underpinnings particularly easy to use. This is in sharp contrast to a book published fifty years ago and

deposited in a library. For digital scholarship, we cannot foresee how maintenance, updates, and migration will work in the future.

[4]"C'est somme toute une sorte de machine à fabriquer des poèmes, mais en nombre limité; il est vrai que ce nombre, quoique limité, fournit de la lecture pour près de deux cents millions d'années (en lisant vingt-quatre heures sure vingt-quatre)" [Queneau 1961, n. p.]

[5]Beckett's short story *Lessness* was first published in French as *Sans*. In his enigmatic story Beckett experimented with random ordering of sentences in the making of fiction.

[6]In a "Reply" to those who commented on his essay, Folsom observed that the *Whitman Archive* is in fact "several databases."

[7]It should be noted that my view of *archive* differs here from that of some commentators. Peter Shillingsburg, for example, remarks that "the level of critical intervention is miniscule in the electronic archive" [Shillingsburg 2006, 156]

[8]In a series of talks in the 1990s John Unsworth and Daniel Pitti began applying the term *thematic research collection* to the type of scholarship under discussion in this paper.

[9]Recent work in archival theory by Heather MacNeill, Elizabeth Yakel, and Michelle Light and Tom Hyry, emphasizes the non-neutral nature of archives themselves and urges the adoption of language such as "archival representation" to highlight the mediating role archivists as they order, interpret, and develop information architectures within socially constructed practice.

[10]Henry Wadsworth Longfellow famously explored the possibility transforming an arsenal into pipe organs of love. See his poem "The Arsenal at Springfield."

[11]I am less concerned that *arsenal* catches on than I am that we recognize the fresh features of new work underway and that we are self-conscious about what we want any new term to convey.

Works Cited

Allen 1963 Allen, Gay Wilson. "Editing the Writings of Walt Whitman: A Million Dollar Project without a Million Dollars." In *Arts and Sciences* 2 (Winter 1963): 8.

Beckett 1969 Beckett, Samuel. *Sans*. Paris : Les Éditions de minuit, 1969.

Beckett 1970 Beckett, Samuel. *Lessness*. London, Calder & Boyars, 1970.

Emerson 1938-1994 Emerson, Ralph Waldo. *The Letters of Ralph Waldo Emerson*. New York: Columbia University Press, 1938-1994.

Folsom 1982 Folsom, Ed. "The Whitman Project: A Review Essay." In *Philological Quarterly* 61 (Fall 1982): 369-394.

Folsom 2007 Folsom, Ed. "Database as Genre: The Epic Transformation of Archives." In *PMLA* 122 (October 2007): 1571-79.

Folsom and Price 1999 Folsom, Ed, and Kenneth M. Price. "The Walt Whitman Archive." Grant Proposal to the National Endowment for the Humanities, 1999.

Higginson 1903 Higginson, Thomas Wentworth. Review of *The Complete Writings of Walt Whitman*. In *The Nation* 76 (May 14, 1903): 400-401.

Hindus 1955 Hindus, Milton. "The Centenary of Leaves of Grass." In *Leaves of Grass: One Hundred Years After*. Palo Alto: Stanford University Press, 1955, 3-21.

Light and Hyry 2002 Light, Michelle and Tom Hyry. "Colophons and Annotations: New Directions for the Finding Aid." *The American Archivist* 65 (Fall / Winter 2002): 216-30.

MacNeil 2005 MacNeil, Heather. "Picking Our Text: Archival Description, Authenticity, and the Archivist as Editor." *The American Archivist*. 68 (Fall / Winter 2005): 264-78.

McGann 2007 McGann, Jerome. "Database, Interface, and Archival Fever." In *PMLA* 122 (October 2007): 1588-92.

Palmer 2004 Palmer, Carole L. "Thematic Research Collections." In *A Companion to Digital Humanities*. Oxford: Blackwell Publishing, 2004, 348-65.

Price 2007 Price, Kenneth M. "Electronic Scholarly Editions." In *A Companion to Digital Literary Studies*. Oxford: Blackwell Publishing, 2007, 434-50.

Queneau 1961 Queneau, Raymond. *Cent mille milliards de poèmes*. [Paris]: Gallimard, 1961.

Reynolds 1995 Reynolds, David S. *Walt Whitman's America: A Cultural Biography*. New York: Knopf, 1995.

Shillingsburg 2006 Shillingsburg, Peter. *From Gutenberg to Google: Electronic Representations of Literary Texts*.

Stauffer 2007 Stauffer, Andrew. "Digital End of the Scholarly Edition." Paper delivered at the Society for Textual Scholarship meeting, New York, March 2007.

Whitman 1978 Whitman, Walt. *Daybooks and Notebooks*. 3 vols. Ed. William White. New York: New York University Press, 1978.

Yakel 2003 Yakel, Elizabeth. "Archival Representation." *Archival Science* 3 (2003): 1-25.