1-25-2018

Increasing Our Vision for 21st-Century Digital Libraries

Elizabeth M. Lorang
University of Nebraska-Lincoln, llorang2@unl.edu

Leen-Kiat Soh
University of Nebraska - Lincoln, lsoh2@unl.edu

Elizabeth Lorang & Leen-Kiat Soh

Opening Keynote, HathiTrust Research Center UnCamp 2018

January 25, 2018

Increasing Our Vision for 21st-Century Digital Libraries

Introduction

Good afternoon, and thank you so much for the invitation to talk with you today and for this opportunity to learn from all of you. It's a pleasure to be here, and we're looking forward to the conversation over the next two days. We are grateful to the event's organizers for entertaining our idea to do a collaborative keynote presentation. Every aspect of this work is a collaborative endeavor, and it felt important to recognize and signal that reality in this, our first keynote, on our work. Thank you for accommodating our request.

We must also begin by acknowledging the other members of our research team, without whom the work of Aida would not be possible. In addition to us, our team members at the University of Nebraska-Lincoln currently include graduate research assistants Yi Liu, Chulwoo (Mike) Pack, and Delaram Rahimighazikalayeh. At the University of Virginia, team members include John O'Brien, Andrew Barrow, and Worthy Martin. We'd also like to acknowledge our advisory board, the members of which have been incredibly generous with their time and their expertise in a variety of areas related to digital libraries: Paul Conway, Jody DeRidder, Adam Farquhar, Emily Gore, Patricia Hswe, Bethany Nowviskie, Ayla Stein, and John Unsworth. Finally, we must acknowledge the generous support of the Institute of Museum and Library Services, as well as the Digging into Data Challenge, both of which currently support aspects of our work; the National Endowment for the Humanities for the initial Start-up grant that allowed us to get off the ground; and the University of Nebraska-Lincoln and University of Virginia.

Next, two caveats:

1) Our presentation today will be taking up, for the most part, a particular type of digital library: digital libraries that are *large-scale* collections or aggregations of *digitized cultural heritage materials*.
2) Our examples and contexts are largely U.S.-, and to a lesser extent U.K., -centric, including in terms of the content that has been digitized and is available via digital libraries as well as in terms of the structures and systems for delivering that content (the digital libraries themselves).

Introduction to Our Work

We've been working together for a couple of years on our project to use digital image analysis and image-based techniques for exploring textual materials, and we presented some early stage work in a summer 2015 article in *D-Lib Magazine*.[1] Since that time, our project and our framing of the project have evolved, as we've moved from thinking about a single, particular challenge—finding a certain type of content, so that we could analyze it—to exploring a range of issues that contribute to the challenge we were originally facing, and to considering how we might intervene in some of those larger issues. Our presentation today will:

1. Read digital library interfaces—or their "main door" interfaces—as glimpses into what we have thus far valued in the development of digital libraries
2. Frame a visual way of thinking about textual materials

[1] Lorang, Elizabeth, Leen-Kiat Soh, Maanas Varma Datla, and Spencer Kulwicki. "Developing an Image-Based Classifier for Detecting Poetic Content in Historic Newspaper Collections." *D-Lib Magazine* 21, no. 7/8 (July 2015). https://doi.org/10.1045/july2015-lorang.

3. Introduce the work of our research team—where we are now, and where we're headed
4. Draw some connections between the parts

This presentation is very much a look into our *thinking in process* and our *work in progress*. At this point, that thinking and work have led us to propose the following ideas:

1. As a community, we can do much more with the digital images we're creating of textual materials than we've heretofore done.
2. We aspire to have additional layers or levels of image analysis become part of the default processing work in the creation of digital libraries, not only as something that happens external or parallel to digital libraries, and not only toward the purpose of generating text.
3. We aspire to more processing up front and iterative processing of materials—so that digital libraries' materials are not "once and done"—and that this "more processing" is presented to users as additional options for how they can explore digital libraries, find materials of relevance, and imagine new possibilities
4. Even as the digital libraries community focuses on supporting computational use of digital libraries—and our research team recognizes that our project very much depends on that computational use being supported—we should not leave behind, in 1998, those users of digital libraries for whom computational use is not their point of entry. (More on that date in a moment.)

Ultimately, we're trying out several ideas, including with the linkages we're drawing and with terminology; you'll see that throughout. And, throughout we're focused not only on the technical challenges—can we achieve a particular result with technology—but also on the social challenges. Our goal is to frame this work holistically. We welcome your feedback and input.

Part 1: 20 Years of Large-Scale Digital Library Development

American Memory

The first version of the Library of Congress's *American Memory* site available via the Internet Archive's Wayback Machine was crawled 20 years ago this month, in January 1998. Most of you are likely familiar with American Memory, but it is the Library of Congress's first effort to digitize millions of items and to make them available online. The effort was launched in 1995, with the goal of the online publication and distribution of more than 5 million items over a 5-year period. The launch of American Memory was an important moment in the development of digital libraries of cultural heritage materials available on the world wide web. The achievement of American Memory required, according to William Arms, "Numerous tricks . . . so that early web browsers could display the materials. They included a specially designed page-turner and a TIFF viewer provided as a plug-in. In an offline batch process, the SGML texts were rendered into HTML, which was stored in the data store. Separate thumbnails were kept for all images." Importantly, as Arms also describes, *American Memory*, "depart[ed] from previous practice [of such emergent digital libraries and] combined access through metadata with full text indexing where possible."[2]

The American Memory website was the public's point of access to the digitized materials. The 1998 version allowed users to conduct a simple keyword search or to click through to a more advanced search where they could limit by collection type, and expand or limit how their search terms should be interpreted (as words, phrase, or variants). Alternately, users might browse collection titles, collection topics, or collection type.

[2] Arms, William Y. "The 1990s: The Formative Years of Digital Libraries." *Library Hi Tech* 30, no. 4 (November 2012): 579–91, p. 5. https://doi.org/10.1108/07378831211285068.

The public site went through a couple of moderate iterations in its early years, until the version of the site launched in October 2004. The fall 2004 version of the website made it easier to browse collections by topic—the topics were featured on the homepage as clickable links, rather than having to be accessed layers in. Buried slightly in the 2004 version was the ability to browse by collections containing particular types of materials, though this was still available, and users could also browse collections by time period and by place. Significantly, however, the site still did not facilitate browsing *items* according to these features, only collections. In sum, from 1998 through 2004, the key additions to how users might find materials within *American Memory* included the added ability to browse by place and time at the collection level, as well as a more straightforward path to browse collections by topic.

The *American Memory* that was available in October 2004 is essentially the same as the American Memory site captured on January 10, 2018--20 years after the first available archived version of *American Memory*. As the orange notice on the current *American Memory* website makes clear, however, items from *American Memory* are now being migrated to elsewhere in the Library of Congress's digital collections architecture. What does the accessibility of *American Memory* materials look like in this new infrastructure, a full 20 years later?

The new portal to Library of Congress Digital Collections, eventually to include all of *American Memory*'s content, continues the main points of entry first presented 20 years ago. The collections remain foregrounded, both as the main browsing unit and as the unit for refining results—by subject, division, and format. The keyword search remains. A key addition of this new environment is the bringing together of *American Memory* with other Library of Congress digital collections—more content, but not necessarily different ways of accessing the content. O.k., but what's the point of this walk down *American Memory* memory lane?

First and foremost, we want to be clear: we are not poking fun at *American Memory*. It is a tremendous resource. We highlight *American Memory* because of its early innovations, because of its 20-year history as a publicly available digital library, and because our ability to access older versions of the site allow us to illustrate quite literally the points we aspire to make. Importantly, we are also using *American Memory* as a stand-in here for most large-scale digital libraries of historic materials.

The early innovations of *American Memory* are now commonplace in the creation and distribution of digital libraries and their content. "Access through metadata with full text indexing" is a basic benchmark that digital libraries of primary source content must meet. Once met, however, the benchmark has remained comfortably close—too close. As a result, the ways most users experience searching and browsing in digital libraries are virtually unchanged in the 20 years since *American Memory* first went online.

Yes, there have been improvements to search algorithms in attempts to minimize the impact of faulty optical character recognition, to account for other words with the same stem as a search query, and to create "smarter" searches in other ways as well. And, implementations of these core functionalities might look somewhat different over time. Holding off for a moment the promise and potential of application programming interfaces, or APIs, however, the ways in which we facilitate finding materials through digital library interfaces look and function effectively—or ineffectively—in the same ways as 20 years ago.

More Product, Less Process

One reason for this stasis, we believe, is that the period of de facto standardization for the design and implementation of digital libraries in the early 2000s dovetails with another

"movement" in libraries and archives, the "more product, less process" philosophy and framework proposed by Mark A. Greene and Dennis Meissner in 2005.

For any in the audience unfamiliar with More Product Less Process, MPLP began as a framework for approaching the processing of physical collections in archives, in order to address the growing backlogs of archival materials that kept many resources largely unknown to researchers. It proposed that the work of arranging, describing, and cataloging collections should first and foremost prioritize getting collections accessible to users. In 2005, Greene and Meissner characterized a MPLP approach as "[describing] materials sufficient to promote use." [3] Putting it another way in 2010, they wrote, MPLP "[Establishes] an acceptable minimum level of work, and [makes] it the processing benchmark."[4]

The MPLP framework specifically addressed the human processing of *physical* collections, but its implications for digitization and digital collections are apparent and significant. Shan C. Sutton described the crossover of MPLP to digitization in 2012: "An ongoing shift away from resource-intensive digitization processes toward large-scale production models is being driven by both MPLP principles and the increasing need to maximize online access to collections in an environment of shrinking staff and budgetary allocations."[5] For digital libraries, MPLP has typically meant:

---

[3] Greene, Mark, and Dennis Meissner. "More Product, Less Process: Revamping Traditional Archival Processing." *The American Archivist* 68, no. 2 (September 1, 2005): 208–63. https://doi.org/10.17723/aarc.68.2.c741823776k65863.

[4] Meissner, Dennis, and Mark A. Greene. "More Application While Less Appreciation: The Adopters and Antagonists of MPLP." *Journal of Archival Organization* 8, no. 3–4 (July 1, 2010): 174–226. https://doi.org/10.1080/15332748.2010.554069.

[5] Sutton, Shan C. "Balancing Boutique-Level Quality and Large-Scale Production: The Impact of 'More Product, Less Process' on Digitization in Archives and Special Collections." *RBM: A Journal of Rare Books, Manuscripts, and Cultural Heritage* 13, 1 (2012): 50-63. https://doi.org/10.5860/rbm.13.1.369.

With regard to digitization and description:

1. Digital image representations
2. Basic bibliographic metadata
3. For printed textual materials, complete electronic text derived via optical character recognition
4. For handwritten and visual materials, basic bibliographic metadata

With regard to Access

1. Text-based searching of bibliographic metadata and complete electronic text
2. Browse interfaces for access via bibliographic metadata

For the most part, these are benchmarks set in 1998.

To be clear, there is much to admire in MPLP values: getting as many materials as possible into the hands of as many users as possible and as quickly as possible sounds ideal. There is, however, tremendous complexity to unpack—how do we determine materials of significance? what types of use should we facilitate or promote, and what types of users do our processing benchmarks privilege or encourage?—and more.

As significantly more materials, and increasingly heterogeneous bodies of materials are digitized and made available, the larger question may be: *What does it really mean to process materials and create digital libraries to an extent that they are findable and usable for researchers?* Are the materials truly accessible? Or, what do the types of accessibility we have privileged suggest about our values? Are the 1998 benchmarks now sufficient for promoting use—and for promoting use equitable to a variety of users?

APIs and Expanded Modes of Access

Some digital libraries of historical materials now offer access to their collections via APIs and other mechanisms, such as bulk downloads. This availability of content via API responds in part to the relatively narrow set of research questions and user experiences that the modern digital library main door, and its existing underlying metadata, make available. Access via API is a significant development for certain types of research and use, such as for analyzing corpora and for creating new collections, tools, and services. The potential is expanded if digital libraries adopt linked data best practices. So, even if "main doors" to digital libraries, such as *American Memory*, structure and frame use in virtually the same way as 20 years ago, access to digital libraries' collections via API and as linked data has certainly opened up additional paths to exploration, analysis, and development.

Just this month, for example, Laura Wrubel from George Washington University Libraries released the "Library of Congress Colors" app, to explore color clusters within LC Digital Collections.[6] Wrubel developed the app alongside the LC Labs team at the Library of Congress, whose mission is to "encourage innovation with Library of Congress digital collections." In a similar vein, the DPLA maintains both its Apps and "For Developers" sections, which encourage remixing and redistribution of materials from the DPLA. The example of "Library of Congress Colors" colors and of apps built onto DPLA collections—and really DPLA itself—are examples of layered digital libraries in practice. Building blocks from one environment are layered with additional information and functionality from others to refine or specialize the digital library's content or services in particular ways, and that additional information and functionality exists separately from the original digital library environment.

[6] http://loc-colors.glitch.me/

What do these newer modes of access and approaches do for our critique of 20 years of stasis with regard to digital libraries' main doors? One view might be that given these other developments, we shouldn't be framing the conversation in terms of "main" and other door positions all, but rather see many points of entry possible. But for better and worse, our digital libraries DO have main doors: hathitrust.org, loc.gov/collections, dp.la, among them. Not only do they introduce visitors to the collections but they highlight or foreground particular types of use and access. They are the portals through which, we would wager, many users, and many types of users, still enter.

Creators of digital libraries must also critique how digital libraries frame access: not only that the materials are available, but that users really are able to locate materials of relevance—and not only *some* example or a great, lucky one—but rather *the most relevant materials to the users' situation and context and the most relevant materials in volume and scale that the researcher requires*.

In our own experiences conducting research and in working with students and faculty alike, the current processing and access benchmarks for digital libraries are not sufficient in this regard. Users cannot find what they need, as studies by Angela Courtney and Harriet Green, and Jody DeRidder and Kathryn Matheny, among others, have also shown.[7] No doubt, education about the resources remains a component, as any academic librarian involved in teaching is well

[7] Green, H. E., and A. Courtney. "Beyond the Scanned Image: A Needs Assessment of Scholarly Uses of Digital Collections." *College & Research Libraries* 76, no. 5 (July 1, 2015): 690–707. https://doi.org/10.5860/crl.76.5.690; DeRidder, Jody L., and Kathryn G. Matheny. "What Do Researchers Need? Feedback On Use of Online Primary Source Materials." *D-Lib Magazine* 20, no. 7/8 (July 2014). https://doi.org/10.1045/july2014-deridder;

aware, but such understanding is not the only barrier to finding relevant materials in digital collections.

To be sure, digital libraries cannot anticipate every type of question and every use case, every situation and context, nor do they need to build for every question and use case. But we can know more about the ways in which current digital libraries fall short for users being to efficiently and effectively find materials of relevance, as well as the strategies they employ. Here, David M. Weigl, Kevin R. Page, Peter Organisciak, and J. Stephen Downie's recent work, "Information-Seeking in Large-Scale Digital Libraries," provides some ideas for the types of question-framing and imagining we might seek to enable and support.[8] The workset model being pursued by HathiTrust for addressing the challenges of access in digital libraries is one potential strategy for approaching these challenges.

We must also acknowledge the power of interfaces and the options they present as "normative" to structure and limit the very types of questions people might even imagine. *What has been the impact of 20 years of keyword search boxes and faceting/browsing on limited textual bibliographic metadata for users' understanding of what's of value, of what's discernible, of where meaning resides in the digitized materials?* Sara Wachter-Boettcher's recent writing about *default settings* in technology seems relevant here, if we understand the default settings of digital library interfaces to be the strategies of the last 20 years. Wachter-Boettcher writes,

[8] Weigl, D. M., K. R. Page, P. Organisciak, and J. S. Downie. "Information-Seeking in Large-Scale Digital Libraries: Strategies for Scholarly Workset Creation." In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 1–4, 2017. https://doi.org/10.1109/JCDL.2017.7991583.

"Defaults . . . affect how we perceive our choices, making us more likely to choose whatever is presented as default . . . ."[9]

Digital libraries have significant power to frame users' perceptions of the materials, and we must take seriously that the choices we make for digital libraries train researchers, for better and worse, to think about the materials in certain ways, to come to ask certain types of questions and not others. Also, that the uses that these interfaces prescribe then come to reinforce the choices we make in digitization, particularly with regard to the extent of processing that seems necessary. Studies of users often indicate that an activity people routinely perform is *keyword searching*, as evidence for the necessity of keyword searching. But of course users use *keyword searching*: not only is it the default option we're provided, sometimes it is the only option. Challenges for digital libraries going forward are both toward getting users access to information and materials of relevance but also to framing the materials in such a way as to honor many sources of information and many paths to and ways of knowing.

Part 2: Digital Images and the Futures of Large-Scale Digital Libraries

Digital Images, Image Analysis, and Visual Meaning-Making

We maintain that creating new experiences and new opportunities for users of digital libraries—particularly those who come through what we are framing as digital libraries' "main doors"—requires additional processing of materials in the creation and distribution of digital libraries. One source for additional processing is the digital images we are creating as we digitize our cultural heritage.

---

[9] Wachter-Boettcher, Sara. *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*. New York, NY: W.W. Norton & Company, Inc., 2017, p. 34.

Our research team is exploring this intersection of issues—the challenges of locating materials of relevance in large collections of digitized historical materials, the levels of processing necessary to support finding materials, and how digital libraries might imagine and create new paths of access within their "main doors"—and the potential of digital images to aid in this work. A major premise of our work is that we do not do nearly enough with the images we are creating in the digitization of the cultural record. In the case of textual materials, they are used for optical character recognition processes, and we post them as digital facsimiles, for those who want to view them further. Information and data we can cultivate from the images may lead to the ability to frame more questions.

At its most basic level, the Aida team's project at this point is to explore the use of image analysis as an approach for aiding content identification, description, and information retrieval in digital libraries and other digitized collections. The majority of existing image analysis work focuses on materials that we understand to be first and foremost visual forms or media: photography, painting, and other of the visual arts, in particular. We, however, are interested in image analysis as a mode for exploring textual materials.

The visual meaning-making of textual materials on its own is not a novel idea. Textual editors and book historians have long noted the power of "bibliographic codes"—material features of texts, including typeface, spacing, and more—to contribute to the meaning that we make of texts. And more recently, they have posited digital environments and modes of analysis for expressing and exploring non-linguistic features of texts.[10] We propose to consider such

[10] McGann, Jerome J. *The Textual Condition*. Princeton Studies in Culture/Power/History. Princeton, N.J: Princeton University Press, 1991; Bornstein, George, "Beyond Codex Editing: A Prototype for the Hypermedia Yeats Project" in Richard F. Finneran, ed., *Yeats: An Annual of Critical and Textual Studies*, vol. 14 (Ann Arbor: University of Michigan Press, 1996), p. 48-58; Audenaert, N., and N. M. Houston. “VisualPage: Towards Large Scale Analysis of Nineteenth-

visual cues as a means toward identification, classification, and exploration—as an additional level of processing that digital libraries might undertake—ultimately to facilitate the ability of users to make connections within digital libraries.

Visual cues can provide powerful clues toward identifying and classifying textual materials or for noticing difference and complexity, similarity and simplicity in textual materials. Such visual meaning making, however, is not captured and maintained in a meaningful way in current default/benchmark processing, nor is it processable for users within the digital library. Most typically, these visual cues are noticeable via browse interfaces but not in such a fashion that enables the user to do more than *observe* them.

At the same time, collection-level descriptions of content type quickly hit the limits of their usefulness, especially for large collections. Even item-level identification can fall short. Letters, for example, might contain poetry, maps, or other drawings, or information in lists that signals a different type of content: still part of the letter, but also a form unto itself. Collections deemed important enough for deep, item-level description might identify the presence of these other forms within a letter, but by and large digital collections do not describe individual items to that degree. And, to the extent that it is so-called "important" collections that get this level of attention, we reestablish their importance and all but guarantee more use for them than for other materials—because people can find more of relevance. Situations such as these are some of the reasons that scholars have worried about the potential of digital libraries, archives, and editions to reinstate dated, problematic literary and cultural canons, rather than build on the recovery and exploding of the canon that has been foundational to literary and cultural studies in particular.

Century Print Culture." In *2013 IEEE International Conference on Big Data*, 9–16, 2013. https://doi.org/10.1109/BigData.2013.6691665.

There is, therefore, real need—pragmatic as well as ethical—in adding to the levels and layers of description in digital libraries and for providing users additional pathways into items that build on these layers of description and extend imagination beyond only browsing and keyword searching. To extend possibilities in this area, we are exploring visual cues, such as those that would be readily apparent in a manuscript or print item that shifts forms and genres, and we are considering the potentials and pitfalls of image-based work for increasing processing and description in digital libraries.

Historic Newspapers

For the time being, we are focused on historic newspapers as a test site for our work. We began our work with digitized newspapers several years ago because of the challenges they pose for researchers. Page segmentation and OCR have been notoriously difficult for historic newspapers, complicating levels of access and types of engagement. Furthermore, newspapers are highly heterogeneous texts, and researchers seek out historic newspapers for many research questions and from many disciplines. The more heterogeneous an item and a digital collection, the more challenges posed for adequate description and for the ability to locate materials of relevance.

Historic newspapers are an example of how a default interface has significantly affected the types of questions researchers might frame and pose, as though all researchers of historic newspapers want to find names or to see issues from a particular date. Many of the questions people wish to pose of historic newspapers are not ones that can be framed in a search box or via newspaper and date browsing. We have treated keyword and concept as analogous, for one, and finding particular types of content has in many ways not improved from pre-digital access, as another example. Therefore, a beginning question for us was whether we could find particular

types of content in a newspaper corpus when that content is not identified as such, using an image-based approach.

In particular, could we find poetic content in historic newspapers? Tens of millions of poems were published in historic newspapers, and the ability to identify these poems and make them more readily findable in a systematic way has profound implications for both teaching and research and for shaping our collective cultural imagination around poetry.

Within most digital collections of historic newspapers, users can:

1. Search for variations on the word poem (poem, poetry, verse, etc.)
2. Search for known lines of poetic content
3. Search for the names of known authors of poetic content

Two-thirds of these options, where you must come in knowing the poem or poet, are contrary to the very idea of having 10s of millions of poems before you and also to how the poems themselves appeared and circulated. These search options are useful in particular cases and for certain types of questions, but they very clearly structure the types of questions one might ask and the poems that will be studied—essentially to those we already know, in some capacity. Meanwhile, searching for variants on the word *poem*, will return only a fraction of the poetry and will exclude verse in many contexts (verse published in death notices, occasional and other poems featured in news stories, poems in advertisements, essentially all of the poems not published in "Poet's Corners," or columns with similar names).

In addition to the visual cues of the newspapers themselves, poems are also highly visual textual forms, presenting a remarkable test case for exploring an image-based approach to the processing of textual materials.

Part 3: The Aida Team's Methods and Approaches to Date

Conceptual Approach

Now, we present an approach to automatically identifying poems in historical newspaper pages. The basic idea stems from how researchers looking for such content, try to visually search through pages and pages of historic newspapers looking for poems—whether the newspapers are original issues in their original format, microphotography reproductions, or digitized versions from originals or microphotography reproductions.

Inspired by the cognitive power of human vision to *abstract what we see*, such as impressions of a poem, and to *learn to recognize patterns associated with the targeted types of texts, such as poems,* over time as a human viewer gets more efficient in identifying poems, we began to model, design, and implement this ability as an automated software methodology. In particular, given a newspaper page, we first carry out segmentation to divide the page into columns, and then into image snippets so that each snippet contains a column and a sufficient number of lines from the newspaper page. Then, we convert each image snippet into a binary image—if the original image was not already binary—identifying the textual pixels and the background pixels. This pre-processing step also involves addressing noise and strengthening textual/object pixels. The binarization step is achieved by assuming a binormal distribution of pixel intensities and identifying the "valley" between two modes.

Once we have the binary image snippets, we proceed to model and capture the visual cues. That is the feature extraction step, with the goal of translating a 2-dimensional image into a vector of numeric values. This step involves going through an image's pixels and performing statistical analyses of the pixel rows and columns. Now, given these vectors of numeric values,

we are ready to build a classifier. In our design, we use the backpropagation artificial neural network approach. It is a supervised learning approach.

Briefly, we identify a set of image snippets where each is expertly labeled to be "poem" or "non-poem" (that is, to contain or not contain poetic content). These labels are considered the ground truth. Then, some of these image snippets are used as the training set. Each image snippet, represented as a vector of numeric values, is fed into the neural network. Based on how the network labels each image snippet and the difference between the label and the ground truth, the network learns to adjust its weights linking its nodes accordingly. At the end of training, the network usually converges to a certain level of success: labeling image snippets correctly when compared to the ground truth. To decide the accuracy of the network, we then apply it to image snippets that have not been used in the training, and compute performance metrics such as precision and recall.

Page Segmentation

Working with the whole pages themselves for feature extraction and classification would create information-overload, because image analysis approaches evaluate each pixel in an image—and each pixel in multiple contexts. Therefore, the first step is page segmentation: dividing up a newspaper page into multiple image snippets. Our current approach relies on first identifying page columns, and then breaking up each page column into multiple overlapping snippets. To identify the page columns, we first convert the newspaper page automatically into a binary image and remove noisy pixels via morphological cleaning to specifically fill in holes or gaps of textual regions, and remove single, isolated textual specks or dots from the background regions. Conceptually, a pixel column of the image is labeled as a page column if that pixel column has a large number of background pixels. Once the page columns are identified, we then

proceed to “segment” or “divide up” each column accordingly. We generate overlapping image snippets to help increase the likelihood of capturing a significant amount of poetic content in a snippet, in relation to the rest of the text.

Pre-Processing, Feature Extraction, and Classification

As alluded to earlier, due to inherent noise in scanned historic newspaper pages, the pre-processing step also involves noise removal and strengthening of object pixels. To remove noise, we blur the original image using 3x3, or 5x5, or 7x7 average filters on each image. After binarization, we apply morphological cleaning that consolidates textual pixels through operations such as erosion and dilation to fill out holes and remove stringy pixels.

To help us capture visual cues, we computed several features extracted from the image snippets. The feature computation methods compute values for three features: left margin whitespace; whitespace between stanzas; and content blocks with jagged right-side edges, which are the result of varying line lengths in poetic content and are in contrast to justified blocks of much newspaper content. At present, an analysis of jaggedness is unique to our project, likely in part because jaggedness emerges as a feature in comparison to the justified prose text. As the first stage of feature extraction, we compute margin statistics, which calculates the mean, standard deviation, and maximum and minimum of the measure of the margin on the left of the image. We include this feature in our algorithm because poetic content typically was typeset with wider left and right margins than non-poetic text. At present, our design focuses on left margins only, since we evaluate qualities of the right side of the poem in relation to jaggedness. The jaggedness algorithm computes the mean, standard deviation, and maximum and minimum measures of the background pixels after the final object pixel in each row. We base our measure of "jaggedness" statistics on the column widths on the right of each image. Finally, we extract

feature attributes that determine the presence of stanzas by looking for whitespace between stanzas. The stanza algorithm computes the mean, standard deviation, and maximum and minimum of the measure of white space between blocks of text. It computes the length of background pixels between paragraphs or stanzas and calculates these attributes. We base our stanza statistics on the measure of the occurrence of spacing between blocks of text.

For the backpropagation neural network, briefly, it consists of three layers: input, middle, and output. Each node of the input layer accepts a parameter or an attribute or an element of the numeric vector extracted for each image snippet. The output node produces a signal that if it is higher than 0.5, then the label is "poem-true," and "poem-false " if it is lower than 0.5. Each edge connecting the layers is weighted, and the weight is adjusted through supervised learning. That is, weights that contribute to a correct label are reinforced further positively, and weights that contribute to an incorrect label are reduced. Each iteration of training involves all image snippets in a training set, and then the percent of correct is computed after each iteration. A network's learning is considered to have converged when the difference in the % of correct between the current iteration and the previous iteration is negligible.

We had 18 input nodes, each for an element of the feature vector of numeric values. We used 9 hidden nodes, and one output node. We used four configurations for the number of training iterations: 1000 iterations, 2000, 3000, and 10,000. We used 16,928 image snippets that were manually, expertly labeled as poem-true or poem-false. The set was balanced: half of the snippets included poetic content, and half of the snippets did not feature poetic content. All these image snippets were derived via our page segmentation technique from *Chronicling America* newspapers from the period 1836-1840. We used an evaluation technique common in machine learning: 10-fold cross validation. We divided the entire set of image snippets into 10 groups.

We trained the network with 9 groups of image snippets and tested the network on the one remaining group. We repeated the process such that each group has been used as the test set.

Results

The average training accuracy for each configuration is around 80%, while the average testing accuracy is around 76%. We also look at two other performance metrics. First, the *precision* metric is basically the % of correct labeling or prediction when an image snippet is labeled to be a poem. The *recall* metric, on the other hand, is the % of image snippets in the set that are identified as poem by the network. In general, we want to have a high precision and a high recall. The average testing precision was around 78%, and the average testing recall was around 73%.

While the results are encouraging, they do show overfitting, where the training results are slightly better than the testing results. Also, the lower recall rate with respect to precision is less desirable: one would want to have the guarantee that most, if not all, image snippets that are "poem" to be identified correctly. We also noticed that the higher number of iterations for training the network failed to significantly improve the training accuracy. For example, when we trained the network with just 1000 iterations, the average training accuracy was about 80%. When we trained the network with 10 times the number of iterations, i.e., 10,000 iterations, the average training accuracy only increased slightly to 81%. This gave us pause: is this visual-cue based, back-propagation neural network approach ultimately viable? Perhaps, the visual cues were not complete, or perhaps the features we identified as humans or how we quantified or extracted the features to approximate the visual cues were not sufficiently effective?

Furthermore, to fully automate the entire end-to-end methodology, there is also another critical step: page segmentation. This step cuts up a newspaper page into snippets, such that each

snippet is a part of one and only one column. We have developed and implemented a relatively successful approach to page segmentation, but there are challenges. The first challenge was caused by low contrast where the newspaper page was over-exposed causing a washed out. The second challenge was caused by significant "bleed through" of the back page, causing the lines of textual pixels to be muddied resulting in "blobs." The third challenge was due to orientation skew and the lack of margin. Our first-generation methodology assumed that the columns are vertical and there were margins, for example, to indicate spaces between columns.

Next Steps and Future Approaches

The insights from the neural network training and classification results, together with the challenges faced in page segmentation, have motivated us to re-think our overall approach. In our 2nd-generation design, we have begun to adopt a connected-component approach to identifying columns and generating image snippets. Instead of looking for spaces between columns, the connected-component approach[11] is a bottom-up approach where textual pixels are connected iteration-by-iteration until all nearby textual pixels are considered as a component. Given these components, boundaries are then drawn, and the subsequent segmentation can then be established to extract individual "rectangles."

As part of our re-thinking, we are also experimenting with convolutional neural networks that have been known to handle inputs with spatial relationships better than backpropagation neural networks. Our first-generation approach was to deal with the visual cues "explicitly"—by modeling, measuring, and converting them into numbers. Presently, we ask the question: Do we

[11] Mitchell, Phillip E., and Hong Yan. "Newspaper layout analysis incorporating connected component separation." *Image and Vision Computing* 22.4 (2004): 307-317.

need to deal with the visual cues explicitly? Or perhaps, are there alternatives to deal with these image snippets without defining and focusing only on explicit visual cues?

Thus presently, we are improving on our original design with better page segmentation enabling a more effective end-to-end solution. This includes investigating the use of convolutional neural networks to obtain an even more generalizable approach to potentially address different types of content as well.

Our preliminary investigation of CNN has been encouraging, and some of our major work over the next several months will be exploring this approach further. We will explore the CNN approach with images derived from our old segmentation technique, as has been the case thus far, as well as on content derived via the new connected-component approach. Likewise, we'll explore our first-generation classification technique on the new connected-component content as well, to explore the range of factors potentially influencing our work.

Extensibility

A major area of work at this time is analyzing and verifying our image analysis approach and extend it so that it is newspaper-agnostic, type-agnostic, and language-agnostic. In addition to Chronicling America newspapers, therefore, we are also working with team members at the University of Virginia to test our approaches on 18th-century British newspapers from the Burney Collection at the British Library. The Burney Collection was one of the first newspaper digitization efforts, and was a partnership between the British Library and Gale. In addition to testing our work on earlier newspapers from a different geographic region, then, we also have an opportunity to explore our approaches as applied to newspapers digitized quite early and within a vendor context. Furthermore, we will be examining how well our approach works on newspapers in languages other than English that are included in Chronicling America—our 1836-1840

exploration already included some Spanish-language papers—as well as in other repositories of digitized newspapers (such as in state newspaper projects). From the outset, we have maintained that a key benefit of looking for visual cues—instead of, or in addition to linguistic cues—is the potential for dealing with multi-language corpora.

Furthermore, to explore whether such an approach can be "type-agnostic" we will be testing this approach on other types of content and genres within historic newspapers. One possibility is to look next at advertisements, because identifying advertisements will be helpful both to those who want to study them, as well as to those who might wish to exclude them from study. Meanwhile, however, in conversation with colleagues at UVA, we've learned that identifying advertisements within 18th-century British newspapers may not be of as much value as looking to other types of content. More valuable might be focusing on trade and commodities information, which would have significance in a variety of contexts and for U.S. and other newspapers as well.

When we've shared about this work, people have no shortage of ideas for the types of newspaper content they want easier access to: recipes, birth and death information, music, weather information. Importantly, these are different genres of interest, and at identified at levels below and in different categories than we've heretofore seen with digital newspaper collections.

While our work focuses on newspapers at this time, depending on the outcomes of this work, we imagine extending it to other print and manuscript forms, both where individual items might be of mixed form and content, as well as across collections. Such an approach is potentially a way of getting toward item-level description when it is the collection itself, as opposed to individual items, that are mixed in form and content. When then coupled with

computer vision approaches to visual items, such as photographs, our approach to image analysis of textual materials may become an even more robust method for processing digitized materials.

As digital artifacts become more and more prevalent, digital libraries become increasingly more voluminous and diverse, threads become longer and connections deeper, effective and efficient access becomes a scalability issue: How can we meta-tag artifacts? How can they be prepared for findability? How can and might one query for artifacts? How can one query for artifacts that one does not yet know exist? How can one look for artifacts that "I would know it when I see it"?

Part 4: Reframing Experiences & Environments

Our emphases and goals have shifted for this work over the last couple of years: When we first partnered up several years ago, the goal was finding poems in historic newspapers, to be able to then do something with the poems—that was what we understood to be our central challenge. Over time, however, we have realized that the challenges that kept a user from finding poetic content in historic newspapers are challenges across many material types, many research domains, and many digital libraries. We have shifted from the goal of creating stand-off resources, whether that's software for identifying the poems or a collection of poems themselves, to exploring *the approach and its implementation* as one that might ultimately become a tool in the toolkit of those creating digital libraries.

In a sense, the shift is in who we understood to be our primary audience: were we creating something that end-users of digital libraries could use, paired with data they accessed via API, or were creating something for those creating digital libraries? Both approaches could be useful, but attempting an intervention at the point at which digital libraries are created and

content digitized—and in the structures for processing materials in digital libraries—seemed like the bigger challenge, with more potential for impact. We credit Trevor Owens, at the time with IMLS and now at the Library of Congress, for helping us to make this shift in our thinking.

Many others have recognized the challenges to findability and of the limiting nature of browsing and keyword searches, and we have seen several approaches to dealing with them. One, as we indicated earlier, has been the opening up of other doors into digital libraries, particularly through bulk access to textual data and metadata through APIs and other strategies. Such opening up has allowed users to develop external tools and services for searching and browsing collections in ways not possible in the primary digital library interface. Another has been to create research environments, in which users might work with the content of digital libraries, using a range of tools developed for text analysis and for studying textual features.

The HathiTrust Research Center is a leader in this regard, and indeed one of the datasets it makes available is Ted Underwood's work on page-level genre identification of volumes in the HathiTrust corpus.[12] We might group these developments under the larger umbrella of "collections as data" approaches, which direct their efforts specifically toward supporting computational use, and typically computational use by the end user. HathiTrust, the Library of Congress, DPLA, and others are leading in this regard, and a recent IMLS-funded project, *Collections as Data*, seeks to help cultural heritage institutions be even more intentional in this work.

We aspire to add to these approaches yet another, one which would see creators of digital libraries increase our benchmarks for digitization and access from the outset, to think afresh about the "levels of description," in Greene and Meissner's words, that are necessary to "promote

[12] https://analytics.hathitrust.org/datasets#genre

use"; about where that description gets expressed and made available; and to think also about to whom it is made available and whose uses are prioritized. Certainly, the outputs of these additional layers and types of processing can be expressed as data and should be made available as such.

But we are first and foremost interested in these additional layers and types of processing for their potential to reshape experiences within what we might still position as digital libraries' "main doors." If many users come to digital libraries via doors that allow them, for all intents and purposes, only the same points of entry and investigation as were possible in 1998, we are letting them down, and we may not be taking seriously enough the power of digital libraries processes and interfaces to shape the questions people can ask and even those that might they imagine. Reframing users' experiences in these environments, and connecting them to materials of relevance, must start from a return to, a reevaluation of, and more process.