NATIONAL ENDOWMENT FOR THE
Humanities

OFFICE OF DIGITAL HUMANITIES

Narrative Section of a Successful Application

The attached document contains the grant narrative and selected portions of a previously funded grant application. It is not intended to serve as a model, but to give you a sense of how a successful application may be crafted. Every successful application is different, and each applicant is urged to prepare a proposal that reflects its unique project and aspirations. Prospective applicants should consult the Office of Digital Humanities program application guidelines at http://www.neh.gov/grants/odh/digital-humanities-start-grants for instructions. Applicants are also strongly encouraged to consult with the NEH Office of Digital Humanities staff well before a grant deadline.

Note: The attachment only contains the grant narrative and selected portions, not the entire funded application. In addition, certain portions may have been redacted to protect the privacy interests of an individual and/or to protect confidential commercial and financial information and/or to protect copyrighted materials.

Project Title: The Visual Page

Institution: University of Houston

Project Directors: University of Houston

Grant Program: Digital Humanities Start-Up Grants, Level 2

NEH Application Cover Sheet
Digital Humanities Start-Up Grants

PROJECT DIRECTOR

Natalie Houston
Associate Professor
236D Roy Cullen Building
Houston, TX 77204-2013
UNITED STATES

E-mail: nhouston@uh.edu
Phone(W): (713) 743-2979
Phone(H):
Fax:

Field of Expertise: Literature - British

INSTITUTION

University of Houston
Houston, TX UNITED STATES

APPLICATION INFORMATION

Title: *The Visual Page*

Grant Period: From 5/2012 to 5/2013
Field of Project: Literature - British
Description of Project: All printed texts convey meaning through both linguistic and graphic signs, but existing tools for computational text analysis focus only on the linguistic content. The Visual Page will develop a prototype application to identify and analyze visual features in digitized Victorian books of poetry, such as margin space, line indentation, and typeface attributes. This will enable scholars to compare documents, identify distinctive or typical books, and track historical changes and influence over very large sets of digitized texts. Current research into such questions is limited by our human capacity to view and compare only a fairly small number of texts at one time. Thus our understanding of their historical significance is based on limited information. Computer analysis can point to significant patterns and trends over a much larger set of texts, which will ultimately transform our understanding of Victorian print culture and the humanities at large.

BUDGET

Outright Request	$49,955.00	**Cost Sharing**	$18,253.00
Matching Request		**Total Budget**	$68,208.00
Total NEH	$49,955.00		

GRANT ADMINISTRATOR

Ms. Denise McGuire
Senior Research Administrator
4800 Calhoun
Houston, TX 77204-2015
UNITED STATES

E-mail: UHPROPOSALS@listserv.uh.edu
Phone(W): (713) 743-9237
Fax: (713) 743-9577

1. Table of Contents

2. List of Participants

Project Director

Houston, Natalie M. (Associate Professor, Department of English, University of Houston)

Technical Director

Audenaert, Neal (Director, Digital Archives Research & Technology Services)

Advisory Board

Drucker, Johanna (Breslauer Professor of Bibliographical Studies, Graduate School of Education & Information Studies, UCLA)
Furuta, Richard (Professor, Department of Computer Science, Texas A&M University; Director, Center for the Study of Digital Libraries; Director, Hypermedia Research Laboratory)
Hughes, Linda K. (Addie Levy Professor of Literature, Department of English, Texas Christian University)
Karadkar, Unmil (Lecturer, School of Information, University of Texas at Austin)
Kooistra, Lorraine Janzen (Professor, Department of English and School of Graduate Studies, Ryerson University)
McGann, Jerome J. (John Stewart Bryan University Professor, Department of English, University of Virginia)
Nowviskie, Bethany (Director, Digital Research and Scholarship, University of Virginia Library)

3. Abstract

All printed texts convey meaning through both linguistic and graphic signs, but existing tools for computational text analysis focus only on the linguistic content. The Visual Page will develop a prototype application to identify and analyze visual features in digitized Victorian books of poetry, such as margin space, line indentation, and typeface attributes. This will enable scholars to compare documents; identify distinctive or typical books; and track historical changes and influence over very large sets of digitized texts. Current research into such questions is limited by our human capacity to view and compare only a fairly small number of texts at one time. Thus our understanding of their historical significance is based on limited information. Computer analysis can point to significant patterns and trends over a much larger set of texts, which will ultimately transform our understanding of Victorian print culture and the humanities at large.

Statement of Innovation

The Visual Page application will allow researchers to pose new questions that engage both the linguistic and graphic signs that jointly convey meaning in printed texts. Research in document image analysis has focused almost entirely on extracting linguistic content. Consequently, the visual components of a page remain invisible to computers. Our work will enable computers to see and analyze many of the graphic features that humanities scholars use to evaluate documents. This will enable large-scale computational analysis of how those features convey meaning.

Statement of Humanities Significance

In books of poetry, the visual appearance of the page (white space, type size, layout) signals and reinforces linguistic features of the text. The Visual Page will provide computational analysis of this visual meaning in digitized page images. Computers can compare much larger sets of texts than humans can in order to identify significant patterns. This tool will enable researchers to gain new understanding of historical changes and significance in book design, contributing to the broader study of literature’s circulation, consumption, and function within Victorian culture.

4. Narrative

The Visual Page

Enhancing the Humanities Through Innovation

The digitization of printed materials published before 1900 has already transformed scholarship in humanities fields such as history and literary studies. Digitized resources improve access for scholars, teachers, students, and general readers. They encourage users of all types to begin asking new kinds of research questions about these texts and their cultural context. Increased access also alters the terms for researching the cultural status and historical significance of particular texts or groups of texts, leading to questions like these asked by Dan Cohen: "Should we be worrying that our scholarship might be anecdotally correct but comprehensively wrong? Is 1 or 10 or 100 or 1000 books an adequate sample to know the Victorians?" Large-scale analysis of digitized materials allows humanities scholars to explore just how unique or representative a particular text or group of texts might be.

How, for example, might the success of Tennyson's 1850 book-length poetic sequence *In Memoriam* be newly understood, if it were compared with all the other books of poetry published in the same year? Most of those books are now available to researchers in digitized form. However, most tools for large-scale digital analysis focus solely on the linguistic content of texts. A researcher could compare the linguistic contents of Tennyson's poem to those of other poems of the same year by tracking word frequency and syntactical clusters within the language of the texts. But if one wished to examine how Tennyson's poem looked on the page in 1850 as compared with other poems, one would be limited to what the human eye can notice and to the constraints of human attention when collating or comparing a large number of texts. To expand beyond those limitations to large-scale analysis, the *Visual Page* project seeks Level II Start-Up Grant funding to develop an open source application to analyze the graphical aspects specific to digitized printed texts.

All printed texts simultaneously convey meaning through both linguistic and graphic signs. As Jerome McGann suggests, "A page of printed or scripted text should thus be understood as a certain kind of graphic interface" (McGann 2001, 199). Words printed on a book's title page, for example, communicate linguistic content (such as the book's title and author's name) that is made more meaningful through the graphic conventions of book publishing. These graphic conventions convey culturally encoded meaning about the importance, audience, and function of the linguistic content on the page or, by extension, of the entire book (Drucker 2009b, 145-64; McGann 1991; McKenzie). The spatial arrangement of text and white space, typeface size and attributes, and the sequencing of the page within the book all combine to signify a title page. It is these conventions, not the linguistic content alone, that distinguishes the book title from the author's name (think, for instance, of George Eliot's *Adam Bede* and other novels whose titles could linguistically signify an author's name). Experienced readers assess, categorize, and evaluate the graphical codes of printed texts quickly, often subconsciously: "we *see* before we read and that . . . predisposes us to reading according to specific graphic codes before we engage with the language of the text" (Drucker 2009b, 242). Graphical aspects of the printed page convey information about the book's historical period, genre, form, cost, audience, function, organization, scope, and design.

The visual elements of printed material can also be deliberately manipulated by their creators for specific effects, as in decorated or illustrated books, or in multimedia works like those by William Blake and Dante Gabriel Rossetti. Such works cannot be adequately represented by their linguistic content alone: "Typographic transcriptions . . . abstract texts from the artifacts in which they are versioned and embodied" (Viscomi 29). Yet the same holds true for all printed texts, not just those that are highly decorated, because printed words themselves function as images: "looking at a set of graphic marks set off by the frame of white space involves the same cognitive processes as would looking at any image" (Mandell 762). Although a full material analysis of a book, including precise page measurements, bibliographic collation, paper watermark and provenance identification, and analysis of binding materials, is not available from the digitized file, digital images of the book's pages offer researchers more information about "the interaction of its physical characteristics with its signifying strategies" than can textual description alone (Hayles 103). Accordingly, most scholarly digital archive projects today recognize the value of this graphical meaning and provide users access to both digitized page images and plain text versions of printed materials. Because the recent digitization projects conducted by research libraries and Google have taken photographic scans of book pages, the graphical meaning of books is available, to varying degrees of fidelity, for a very large corpus of digitized items.

For the Start-Up Grant period, we will be working with a data set of 300 books of poetry (approximately 60,000 images) published between 1860-1880. In books of poetry printed after 1800, the visual appearance of the page often signals and reinforces linguistic features of the poem. The relative amount and distribution of white space on the page cues the reader to the presence of poetry and even to specific verse forms such as the sonnet, which was frequently surrounded by ample marginal space. The graphic conventions of line capitalization, punctuation, and indentation visually distinguish many kinds of poetry from prose. Extra leading, or white space, between poetic stanzas and the indentation of poetic lines reinforce rhyme patterns and formal structures of historically specific verse forms.

The *Visual Page* project seeks Level II Start-Up Grant funding to develop a prototype application to identify and analyze visual features in digitized Victorian books of poetry, such as margin space, line indentation, ornamentation, and text density. The proposed application will integrate tools for machine learning (i.e., discovering which features help to visually distinguish books from two different publishers), pattern analysis and classification (identifying groups of visually similar works or finding other poems that look like Tennyson's) and visualizing relationships between poems (juxtaposing sets of images or scatter plots based on computational measures of visual similarity or difference).[1]

We anticipate that computational analysis of these visual features will reveal new ways of thinking about both the printed book and its digitized forms. Because scholarship in the history of printing, publishing, and book design is grounded in empirical analysis of material artifacts, one of the important goals of the Start-Up Grant period will be to demonstrate the validity of this computational analysis. We will conduct bibliographical measurements of a sampling of the books contained in the digital data set to verify the measurements and comparisons generated by the *Visual Page*.

Much of the previous scholarship on Victorian publishing practices and page design has focused on particular publishers, illustrated books, or particular authors, such as William Morris or Oscar Wilde, whose highly decorated books represented particular political or aesthetic goals and strategies.[2] The *Visual Page* application will enable researchers to examine the graphical aspects of these decorated books as well as those of ordinary books of poetry, thereby contributing to a broader understanding of literature's circulation, consumption, and function within Victorian culture.

The *Visual Page* project will be valuable to researchers in the fields of literature, history, cultural studies, and media studies. This application will enable researchers to expand the scope of current research questions about the material book to a larger scale. For example, this application will allow researchers to explore:

<u>Similarities and differences between sets of printed materials</u>. How do books published by Macmillan and those by Bell and Daldy during the second half of the nineteenth century differ in their visual appearance? How do the text pages of illustrated books of poems compare with those in books

[1] See appendix for technical details.

[2] See, for example, Frankel, Helsinger, and Kooistra.

without illustration? How do these differences correlate with other features of these texts, such as price, distribution networks, poetic forms, or themes?

Historical changes in printed materials. When do ornamental initial capital letters become widely used in books of Victorian poetry? When and how do they arise, change, or disappear? How does their usage correlate with specific publishers, authors, or poetic forms?

Measuring and identifying distinctive features and/or distinctive books. What are the most unusual books of poetry published in the 1860s? What makes them different from other books in their visual appearance? Does that difference correlate with specific authors, publishers, poetic forms, or themes?

Measuring and identifying representative or typical books. What does an average or typical book of poetry look like in the 1860s? What might this suggest about Victorian reading practices?

Influence and imitation in the design of printed materials. How were distinctive book designs by Joseph Cundall or William Morris imitated by others? Do these artistic designs have any effect on mass book publishing?

Although these research questions are specific to the set of Victorian books of poetry we will be using for this Start-Up Grant, similar questions are of interest to scholars working on other kinds of printed materials and in other periods.

The extent to which we can currently research such questions is constrained by our human capacity to view, compare, and understand only a limited number of texts at one time. Thus our understanding of what constitutes a significant or representative text is based on relatively limited historical information. Computational analysis can point to significant patterns and trends over a much larger set of texts, which will lead humanities researchers to study previously unknown texts as well as to understand canonical texts in new ways. Ultimately, such large-scale research will transform the boundaries and definitions of humanities research by changing our understanding of key ideas, developments, and conflicts within print culture.

Environmental Scan

From the shift from oral reading to silent reading made possible by the introduction of word spacing in the ninth century to the study of the relation of text and image in twenty-first century graphic novels, humanities scholars in many fields and periods have an interest in the graphical aspects of literary and nonliterary texts.[3] Specialized studies in material textuality occur within the areas of book history, print culture, bibliography, and media studies.[4] In addition, scholars in literature and history frequently rely upon the visual aspects of printed texts either explicitly or implicitly in selecting texts as distinctive or representative examples of historical, social, artistic, or political trends, patterns, or tendencies.[5]

Today, scanned document images constitute the focus of many scholarly editions and digital archives. Most scholarly archives of print materials, like those for Dante Gabriel Rossetti, Walt Whitman, or Shakespeare, provide page images of digitized books as well as plain-text transcriptions.[6] Document imaging technology is also used to preserve fragile records from the ancient world, as in the Archimedes

[3] See, for example, Benton, Dowling, Enenkel, Hoagwood, Johns, Levenston, McGill. Ong, Raven, and Saenger.

[4] See, for example, Darnton, Erickson, Frankel, Kooistra, Manguel, Maruca, St. Clair, Sher, and Smith.

[5] See, for example, Anderson, Calè, Chartier, Eisenstein, and Hilliard.

[6] http://www.rossettiarchive.org; http://www.whitmanarchive.org; http://www.quartos.org

Palimpsest, Duke University Papyrus Archive, and the Homer MultiText Project.[7] Humanities researchers in a variety of fields have come to expect access to document images within digital archives and exhibits. By providing new tools for analyzing scanned document images, the *Visual Page* will build upon this existing body of scholarship and contribute to new research with these materials.

Computational text analysis tools are now available to researchers through the *Text Analysis Portal for Research* (TAPoR) and the *Software Environment for the Advancement of Scholarly Research* (SEASR).[8] The *Metadata Offer New Knowledge* (MONK) project offers text analysis within a defined set of literary texts.[9] A wide variety of commercial text analysis software tools are also available. These tools draw upon work in pattern recognition and machine learning (Bishop, Mitchell).

A number of text analysis research projects compare specific linguistic units across large data sets of nineteenth-century materials. At the Stanford Literary Lab, current research examines the stylistics of the novelistic sentence, comparing syntactical structures in the text of 250 nineteenth-century British novels.[10] The *Wordseer* project analyzes grammatical structures within nineteenth-century slave narratives.[11] The *Reframing the Victorians* project is text mining 1,681,161 English books published between 1789-1914 to examine patterns in book titles, word usage, and Biblical citation.[12] These valuable large-scale analyses of nineteenth-century printed materials focus solely on their linguistic content. The *Visual Page* application will provide tools to expand such analyses into the visual or graphic content of these materials.

Computational analysis of document images has been a field of ongoing, intensive research since the late 1960s (Nagy 2000). Of particular relevance to the *Visual Page* is work in the sub-field of document layout analysis (Mao et al.). Layout analysis is typically approached as a pre-processing step for optical character recognition that seeks to decompose the page image into regions of homogeneous text that can be sent to the OCR engine for recognition. The *Visual Page* will draw on these techniques to build a model of the page from which visual features can be extracted.

Research into data visualization examines how the presentation of information in visual forms can assist researchers in analyzing and understanding their material. Useful examples of visualization tools and practices can be found at projects such as *Visual Complexity*, *Flowing Data*, and *Many Eyes*.[13] We will be evaluating some of these tools for possible use in the presentation of research results from the *Visual Page*.

Lev Manovich's research group, Software Studies Initiative, has recently released a new open source software tool, *ImagePlot*, for displaying large sets of visual information so as to enable human perception to see analytical patterns within the data.[14] *ImagePlot* produces scatterplots or line graphs that display thumbnail images from the dataset as the data points. Graphs can be produced using tagged metadata or image features such as hue, saturation, and brightness as the axes for comparison.

What distinguishes the *Visual Page* application from *ImagePlot* is our specific focus on semantically rich visual features specific to printed texts. These include features that may be implicit in

[7] http://www.archimedespalimpsest.org; http://scriptorium.lib.duke.edu/papyrus; http://chs.harvard.edu/chs/homer_multitext

[8] http://portal.tapor.ca; http://seasr.org

[9] http://www.monkproject.org

[10] http://litlab.stanford.edu

[11] http://www.eecs.berkeley.edu/~aditi/projects/wordseer.html

[12] http://www.dancohen.org/2010/10/04/searching-for-the-victorians

[13] http://www.visualcomplexity.com/vc/; http://flowingdata.com; http://www-958.ibm.com/software/data/cognos/manyeyes/

[14] http://lab.softwarestudies.com/p/imageplot.html

the page design (such as density of text pixels to background pixels on the page) as well as explicitly designed features such as ornamental capitals and line indentation. Secondly, our application utilizes computational analysis over these visual features in order to identify meaningful relationships. Many of these meaningful relationships within the data set may not be apparent through visualization alone.

History and Duration of the Project

Project Director Natalie Houston has a long-standing research interest in the design and cultural significance of printed texts and her publications include articles on the design of Victorian sonnet anthologies and the appearance of poetry in Victorian newspapers (Houston 1999, 2008). In 2009, as preliminary research for this project, she began using *ImageJ* image analysis software to analyze digitized page images of Victorian books of poetry, using percentages of pixels in binarized images as a measurement of text density on the page. She also used image-editing tools to create multiple visualizations of page images, such as reduced image tiling and translucent palimpsest layering, to reveal trends in page layout and relative text block dimensions. She presented this research at the Society for the History of Authorship, Reading, and Publishing (SHARP) conference in June 2009 and at the North American Victorian Studies Association (NAVSA) meeting in November 2010. In 2011 she began developing a database of digitized files and metadata for books of Victorian poetry published 1860-1880 that will serve as the data set for this Start-Up grant.

Technical Director Neal Audenaert has worked on several NSF and NEH funded projects that focus on image-based collections. These projects include: the Cervantes Project (Urbina, et al. 2006), the Online Picasso Project (Audenaert, et al. 2007), and the Nautical Archaeology Digital Library (Audenaert and Furuta 2009). His dissertation focused on understanding how humanities scholars use visually complex source material (including both physical documents and digital facsimiles) (Audenaert and Furuta 2010) and designing an interactive visual environment to facilitate both sense-making tasks and in depth analysis of these documents (Audenaert, et al. 2010). This work is part of a broader, NSF-funded project that investigates how advanced computational techniques can enable people to distinguish and manipulate the structural components of digital facsimiles so as to deconstruct a document into its constituent parts for analysis.

Future development of the *Visual Page* project will explore its applicability to digitized printed texts in different genres and from different historical periods. In addition, we plan to develop additional user interface tools to assist humanities researchers in understanding the visual analysis this project offers. When matured, the application developed in the *Visual Page* will be made available for researchers using publicly available digitized materials as well as to digital archives wishing to integrate visual analysis into their project sites. We expect this application will deepen research in existing archives as well as lead to the creation of new research projects, exhibits, and digital collections.

Work Plan

We are requesting Level II Start-Up funds to support one year of work (May 2012- May 2013) resulting in a proof of concept application. Major tasks will include:

Data Preparation (May-Aug 2012): Collect PDF files of digitized books; divide them into separate page images; create book, section, and poem metadata; identify key dimensions to be analyzed (such as margin size, header placement, spacing between stanzas, etc).
Feature Extraction (Aug-Dec 2012): Develop feature extraction process; conduct ground truth analysis; evaluate the accuracy of digital feature extraction with bibliographic measurements of a sampling of physical books.
Pattern Recognition (Jan-March 2013): Conduct classification and clustering analysis.

Final Results and Presentation (Mar-May 2013): Develop visualization plan for research results, using existing open-source tools; writing and presentation of research results.

We will develop the feature extraction and pattern classification algorithms in C++ and Python. The user interface components will be developed as JavaScript powered rich client applications using Processing.js to support visualization. Throughout the grant period, we will maintain an ongoing blog which will document and critically engage with both the technical and theoretical developments and discoveries of the project. With successful development of a prototype application during the Start-Up Grant period, we will seek further grant funding to support the development of a mature application.

Staff

Project Director Natalie Houston, Associate Professor of English at the University of Houston, will be responsible for: general project design and management; feature identification; data preparation; bibliographic verification; project website creation and management; and writing of articles and presentations about project results.

Technical Director Neal Audenaert, director of Digital Archives Research & Technology Services, will be responsible for software design and development; feature extraction algorithms; pattern classification algorithm selection and application; application interface design and development; and writing of articles and presentations about project results.

Final Product and Dissemination

The prototype *Visual Page* application will be implemented as a suite of visual analysis tools. This will enable someone with programming ability to add new feature identifiers or pattern classification algorithms or to re-purpose existing components to meet new needs. This will allow the application to incorporate technological advances in the field of image analysis and to support scholarly innovation as humanists identify new visual features that they would like to investigate. The application will be made available through the source code repository GitHub and disseminated under the terms of the Apache 2.0 License.[15]

In addition to the project blog described above, research reports and the project white paper will be made available from the project website. Both Natalie Houston and Neal Audenaert will propose papers based on this project for presentation at scholarly conferences devoted to Victorian studies, digital humanities, image analysis, and machine learning.

[15] https://github.com/; http://www.apache.org/licenses/LICENSE-2.0.html

NATIONAL ENDOWMENT FOR THE HUMANITIES

Sample Budget Form (rev. 6/2011)

Applicant Institution: University of Houston
Project Director: Natalie Houston
Project Grant Period: 5/1/2012 - 5/30/2013

See online Budget Instructions (4-page PDF)

	Computational Details/Notes	**(notes)**	**Year 1**	**(notes)**	**Year 2**	**(notes)**	**Year 3**	**Project Total**
			5/1/2012 - 9/30/2013					
1. Salaries & Wages								
Project Director Natalie Houston	Academic year salary: Ex. B6 (yr1)	23%	Ex. B6					Ex. B6
								$0
2. Fringe Benefits								
Project Director Natalie Houston	FB based on actuals using the UH FB calculator		Ex. B6					Ex. B6
								$0
3. Consultant Fees								
								$0
4. Travel								
Project Director Natalie Houston	1-day Meeting at the NEH Offices in Washington DC.		$1,000					**$1,000**

	Computational Details/Notes	(notes)	Year 1	(notes)	Year 2	(notes)	Year 3	Project Total
5. Supplies & Materials								
								$0
6. Services	**Subcontract**							
Technical Director Neal Audenaert			Ex. B6					Ex. B6
7. Other Costs								
								$0
8. Total Direct Costs	**Per Year**		**$47,280**		**$0**		**$0**	**$47,280**
9. Total Indirect Costs	**Per Year**		$20,928					**$20,928**
Indirect Cost Calculation: a. Rate: 49.5% of MTDC b. Federal Agency: DHHS c. Date of Agreement: 08/24/11								
10. Total Project Costs (Direct and Indirect costs for entire project)			$68,208					**$68,208**
11 Project Funding								**$49,955**
a. Requested from NEH	Outright:		$49,955					**$49,955**
	Matching Funds:							
	Total Requested from NEH:		$49,955					**$49,955**

	Computational Details/Notes	**(notes)**	**Year 1**	**(notes)**	**Year 2**	**(notes)**	**Year 3**	**Project Total**
b. Cost Sharing	Applicant's Contributions:		$18,253					**$18,253**
	Third Party Contributions:							
	Project Income:							**$0**
	Other Federal Agencies:							**$0**
	Total Cost Share:		$18,253					**$18,253**
12. Total Project Funding			$68,208					**$68,208**

Budget Narrative

Project Director Natalie Houston will be spending 23% of her time and effort during the grant period towards this project, for a value of Ex. B6. 6% of her time and effort will be charged to NEH project funds and the remainder will be provided through cost-sharing from the University of Houston.

Technical Director Neal Audenaert will be contracted for 240 hours of work for the project for a total of Ex. B6. His itemized time is as follows:

	hours	hourly cost	total
Data Preparation (7 May-31 Aug 2012)	40	Ex. B6	Ex. B6
Feature Extraction (1 Sept-28 Dec 2012)	80	Ex. B6	Ex. B6
Pattern Recognition (2 Jan 2013-29 Mar 2013)	80	Ex. B6	Ex. B6
Dissemination of Results (1 Apr 2013-31 May 2013)	40	Ex. B6	Ex. B6
TOTAL	**240**		Ex. B6

6. Biographies

Project Director Natalie M. Houston is an Associate Professor of English at the University of Houston, where she has taught since 1998. Her research on Victorian print culture, women poets, poetry anthologies, and the Victorian sonnet, has appeared in journals such as *Victorian Studies, Victorian Poetry*, *Yale Journal of Criticism*, *Romanticism and Victorianism on the Net, Essays and Studies*, and *Studies in the Literary Imagination,* as well as in *The Blackwell Companion to Victorian Poetry*. She also prepared a critical edition of Mary Elizabeth Braddon's novel *Lady Audley's Secret* (Broadview Press 2003). She is currently a Co-Project Director and Technical Director for the *Periodical Poetry Index*, a research database of English-language poems published in nineteenth-century periodicals (public launch expected December 2011). She has presented initial research for the *Visual Page* in two recent conference papers: “Reading White Space: the Visual Codes of British Poetry of the 1860s” at the Society for the History of Authorship, Reading, and Publishing (SHARP) (University of Toronto, June 2009) and “Methodology, Scale, and the Digital: Analyzing Victorian Poetry’s Visual Codes” at the North American Victorian Studies Association (NAVSA) (University of Montreal, November 2010).

Technical Director Neal Audenaert defended his dissertation in Computer Science and Engineering at Texas A&M University in September 2011 (Ph.D. to be awarded in December). Since 2003, he has worked at the TEES Center for the Study of Digital Libraries, where he has worked on a number of successful digital humanities projects including the Cervantes Project, Digital Donne, the Picasso Project and the Nautical Archaeology Digital Library. His own research focuses on understanding scholarly research as a creative process and designing interactive systems to support scholars’ work. He is currently working with document image analysis techniques in order to make the visual structure of documents more accessible to both users and computational analysis. Neal has published 14 peer reviewed articles and conference papers in venues such as the ACM/IEEE Joint Conference on Digital Libraries, the European Conference on Digital Libraries, *Digital Humanities*, and *Literary and Linguistic Computing*. He serves as the continuing issues editor of the *Bulletin* of the IEEE Technical Committee on Digital Libraries and is a member of the Digital Humanities Working Group at Texas A&M University. Neal currently directs Digital Archives, Research & Technology Services (DARTS), a non-profit software consulting firm that provides technical assistance to design and implement digital tools to support scholarship in the humanities.

Advisory Board members will receive quarterly updates on research methods and findings for *The Visual Page*. They have agreed to provide written or oral comments to the project participants then or at other intervals during the project. There will be no formal meeting convened of the project’s Advisory Board during the Start-Up grant period. Advisory Board members include:

Johanna Drucker, Breslauer Professor of Bibliographical Studies, Graduate School of Education & Information Studies, UCLA. Selected publications include: *SpecLab: Digital Aesthetics and Speculative Computing* (2009), *Graphic Design History: A Critical Approach* (2008); *The Visible Word: Experimental Typography and Modern Art* (2004); *Figuring the Word: Essays on Books, Writing, and Visual Poetics* (1998).

Richard Furuta, Professor, Department of Computer Science, Texas A&M University; Director, Center for the Study of Digital Libraries; Director, Hypermedia Research Laboratory. Publication list available at: http://www.csdl.tamu.edu/~furuta/tamubio.pdf

Linda Hughes, Addie Levy Professor of Literature, Department of English, Texas Christian University. Selected publications include: *The Cambridge Introduction to Victorian Poetry* (2010); *Graham R: Rosamund Marriott Watson, Woman of Letters* (2005); *Victorian Publishing and Mrs. Gaskell's Work* (with M. Lund, 1999); *The Victorian Serial* (with M. Lund, 1991).

Unmil Karadkar, Lecturer, School of Information, University of Texas at Austin; Ph.D. candidate, Texas A&M University. Publication list available at: http://www.csdl.tamu.edu/~unmil/

Lorraine Janzen Kooistra, Professor, Department of English and School of Graduate Studies, Ryerson University. Selected publications include: *Poetry, Pictures, and Popular Publishing: The Illustrated Gift Book and Victorian Visual Culture* (2011); *Christina Rossetti and Illustration: A Publishing History* (2002); *The Artist as Critic: Bitextuality in Fin-de-Siècle Illustrated Books* (1995). Co-editor of *The Yellow Nineties Online* and *The Children's Literature Archive*.

Jerome J. McGann, John Stewart Bryan University Professor, Department of English, University of Virginia. Selected publications include: *The Point is to Change It: Poetry and Criticism in the Continuing Present* (2007); *Radiant Textuality: Literature Since the World Wide Web* (2001); *Black Riders: The Visible Language of Modernism* (1993); *The Textual Condition* (1991). General Editor, *The Rossetti Archive*.

Bethany Nowviskie, Director, Digital Research and Scholarship, University of Virginia Library. Co-PI, *Omeka + Neatline* (Library of Congress); PI, *Institute for Enabling Geospatial Scholarship* (NEH). Senior Advisor, NINES. Created NINES Collex software and as Design Editor, Rossetti Archive, created initial and present user interface.

7. Data Management Plan

Expected data. The primary data to be produced, maintained and distributed by the *Visual Page* is the software we create. This will consist of C++, Python and Java source code, configuration files, generated source code documentation, and higher-level system and user documentation. We will refer to this collectively as the software. The software will be developed as a publicly available open source project from the outset. We will use GitHub as our source code repository and version control system.

In addition to the software, we will also collect digital facsimiles of books of poetry to use in developing and testing our application. This collection will, as permitted by copyright law and licensing restrictions, be made available to the public for download via our project web site. Where copyright law or licensing restrictions prevent such distribution, we will provide sufficient detail to enable those who are interested in replicating our work to locate and retrieve these documents. We will maintain metadata that we associate with these documents (e.g., author, publisher, publication date) internally using a relational database. This metadata will be made publicly available as a CSV file available for download at our web site. Internally, we may use a number of derived forms including high-resolution and thumbnail page images along with metadata to link these derived forms to their sources. Instead of distributing this derived data, we will provide the tools required to create it along with detailed instructions for doing so.

Throughout the project we will create and use data files that describe the visual features extracted from the documents. To enable analysis and evaluation of our work, we will prepare public versions of these data sets to be distributed under the Creative Commons Attribution 3.0 license. These data sets will be available from the *Visual Page* project web site.

Period of data retention*.* All software and data will be made accessible throughout the course of the project as it is developed. The visual feature data-sets will be prepared and made available once we anticipate that no more work will be performed to extract visual features from the documents. Upon completion of the project, the web site and all related data will be transferred to servers maintained by Natalie Houston's institution or another hosting provider. The project's software will be maintained on GitHub by DARTS for a period of not less than 5 years.

Data formats and dissemination. The poetry books will be stored and disseminated as PDF documents, compressed as a gzipped tar file for convenience. Data sets consisting of extracted visual features will be stored and disseminated as ARFF (Attribute-Relation File Format) and CSV (Comma Separated Values) files. All resources will be made publicly available. With the exception of the digital facsimiles for poetry books, we do not anticipate any privacy, confidentiality, security, intellectual property or other rights or requirements that will impact our ability to store and dissemniate this data. For digital facsimiles that we are not permitted to re-distribute, we will provide documentation for where we obtained our sources. All digital copies of restricted data will be destroyed upon completion of the project.

The software will be distributed as source code under the terms of the Apache 2.0 and the documentation will be distributed under the terms of the Creative Commons Attribution 3.0 license. Within the scope of the Start-Up phase of this project, we do not anticipate distributing compiled, executable binaries of our software.

Data storage and preservation of access. All data (excluding software) will be stored and managed for the duration of the project on a sever provisioned by Digital Archives, Research & Technology Services (DARTS) using Amazon Web Service (AWS). Data will be stored on an Elastic Block Store (EBS) device which will be automatically backed up on a weekly basis to AWS Simple Storage Solution (S3). All software created for the project will be stored using GitHub and accessible through the Git source code control application and via the GitHub web site.

NINES

http://nines.org

September 25, 2011

To Whom It May Concern:

As Associate Director of NINES, I write to recommend that an NEH Digital Humanities Start-Up Grant be awarded to Natalie Houston for “The Visual Page” in order to create a means for digitally analyzing Victorian books of poetry. Whitespace and typographical arrangements proffer meanings to which we have become too habituated to fully grasp. I know from working with literary annuals that even the margins of pages suggest a book’s cost and typical readership: what other kinds of meanings are proffered by the look of the page?

We need at this juncture to be building tools for discovery, and the Visual Page is perfect given the sheer number of page images available for analysis as well as our as yet limited knowledge about how cultural capital is communicated to the eye. Moreover, it provides empirical evidence for what has been heretofore only anecdotal information.

Though the process is being performed on Victorian poetry, it will provide a model for researchers to analyze numerous data sets: EEBO and ECCO, for instance, which offer us collectively about 400,000 texts of page images, as well as everything in Google books. The Visual Page project will show us what we can do with page images besides read them: it offers an eye to detail that can work across massive data sets.

I can vouch as well for the consummate work of Neal Audenaert who received his Ph.D. recently from the Center for the Study of Digital Libraries here at Texas A&M. This is a worthy and well-planned project.

Sincerely,

Laura Mandell

Laura Mandell
Associate Director, NINES
Director, Initiative for Digital Humanities, Media, and Culture
Professor of English, Texas A&M University

DIGITAL ARCHIVES RESEARCH & TECHNOLOGY SERVICES (DARTS)

Project Agreement

Project Title	Visual Page		
Customer	**Natalie Houston**	**DARTS Lead**	**Neal Audenaert**
E-mail	nhouston@uh.edu	E-mail	Ex. B6
Phone	713-743-2979	Phone:	Ex. B6
Mailing	Department of English University of Houston Houston, TX 77204-3013	Mailing:	Ex. B6
Start Date:	7 May 2012	End Date:	31 May 2013

GENERAL TERMS & CONDITIONS

Project Detail

This agreement covers software development work to be performed by DARTS in support of the Visual Page project, directed by Natalie Houston and detailed in the attached NEH proposal. This agreement is conditioned upon successful funding of the proposed project.

Billing

This project will be billed on an hourly basis at our standard rate of Ex. B6 hour, not to exceed 240 hours. We will send invoices for completed work to be paid within 15 working days.

Quality Assurance

Natalie Houston will be responsible for evaluating all work (with the assistance of the project's advisory board) to ensure its scholarly merit and that the work performed by DARTS staff conform to accepted professional standards.

Neal Audenaert will be responsible to ensure that all technical work and related documentation conforms to accepted professional standards (given the project status as a proof-of-concept prototype). He will be responsible for selecting appropriate algorithms and/or adapting existing tools in order to effectively meet the scholarly objectives of the project as set forth by Natalie Houston.

Project Cancellation

Once funded, this project may be terminated by Natalie Houston, if, at her sole discretion, DARTS technical services do not meet her expectations or if it appears that DARTS will not be able to make reasonable progress to ensure the scholarly contribution of the project. Upon being notified of cancelation DARTS will cease all work, and turn over all software and document produced to date to Natalie Houston pending final payment for completed work.

In order to ensure that it can meet its commitments to other clients, DARTS may cancel the project if failures to act on Natalie Houston's part result in excessive delays to the mutually agreed upon schedule (more than 30 days). DARTS must notify Natalie Houston of the failure 15 business days prior to any such cancelation and provide reasonable steps to that can be taken to the project to a mutually agreeable schedule.

SCOPE OF WORK

DARTS will provide software design and development services in support of the Visual Page project. This work will include

- Tools to divide source PDF documents into separate page images; create book, section, and poem metadata
- Develop feature extraction process; assist in conducting ground truth analysis to evaluate accuracy
- Design and implement machine learning algorithms to facilitate document clustering and classification
- Assist in development of data visualization plan

DARTS will customize existing page layout analysis algorithms and/or implement algorithms previously reported in the research literature, tailoring these tools as needed to meet the specific requirements of the project. We will also implement machine learning techniques to identify relationships and patterns in the document images based on these extracted features.

All software developed will be stored and disseminated via GitHub under the terms of an Apache 2.0 license. DARTS will retain copyright ownership of all developed software.

This work will be performed with the understanding that we are developing a proof of concept system intended to demonstrate the key concepts set forth by Natalie Houston in the attached NEH proposal. The specific feature identification algorithms and machine learning techniques to be implemented will be selected by mutual agreement between DARTS and Natalie Houston based on research priorities and available funding.

PROJECT ROLES AND RESPONSIBILITIES

Neal Audenaert DARTS President & Executive Director

Neal Audenaert will represent DARTS and will perform and/or supervise for software design and development; feature extraction algorithms; pattern classification machine learning algorithm selection and application; application interface design and development; and technical documentation for the project.

Natalie Houston Associate Professor, University of Houston; PI of Visual Page Project

Natalie Houston, as principal investigator for the project will be responsible for general project design and management; selection of visual feature to be recognized; data preparation; bibliographic verification; project website creation and management; and writing of articles and presentations about project results.

PROJECT MILESTONES

Milestone	Description	Target Date
Data Preparation	Collect PDF files of digitized books; divide them into separate page images; create book, section, and poem metadata; identify key dimensions to be analyzed (such as margin size, header placement, stanza spacing, etc).	31 August 2012
Feature Extraction	Develop feature extraction process; conduct ground truth analysis; evaluate the accuracy of digital feature extraction with bibliographic measurements of a sampling of physical books	28 December 2012
Pattern Recognition	Design, implement and conduct classification and clustering analysis.	29 March 2013
Dissemination	Develop visualization plan for research results, using existing open-source tools; writing and presentation of research results.	31 May 2013

ESTIMATED BUDGET (DARTS)

Work Item	Hours	Cost	Total
Data Preparation	40	Ex. B6	Ex. B6
Feature Extraction	80		
Pattern Recognition	80		
Dissemination	40		
TOTAL:	**240**		$

Signatures

Neal Audenaert
President & Executive Director, DARTS

Natalie Houston
Associate Professor, University of Houston

Letter of Commitment to join the Advisory Board for *The Visual Page*

I hereby agree to participate as a member of the Advisory Board for *The Visual Page*, a Start-Up Grant project being proposed to the NEH Office of Digital Humanities by Natalie M. Houston (Associate Professor, Department of English, University of Houston).

If funded, the term of the project and my participation in its Advisory Board will be from May 2012 – May 2013. There will be no remuneration for my participation on this Board.

As a member of the Advisory Board, I will receive quarterly updates on the project's research. I agree to provide comments to the project participants then or at other intervals during the project. Comments may be written or oral. There will be no formal meeting convened of the project's Advisory Board.

____
Johanna Drucker

Breslauer Professor, UCLA, Department of Information Studies
____
Title and Institutional Affiliation

_______September 23, 2011_______________________________________
Date

Letter of Commitment to join the Advisory Board for *The Visual Page*

I hereby agree to participate as a member of the Advisory Board for *The Visual Page*, a Start-Up Grant project being proposed to the NEH Office of Digital Humanities by Natalie M. Houston (Associate Professor, Department of English, University of Houston).

If funded, the term of the project and my participation in its Advisory Board will be from May 2012 – May 2013. There will be no remuneration for my participation on this Board.

As a member of the Advisory Board, I will receive quarterly updates on the project's research. I agree to provide comments to the project participants then or at other intervals during the project. Comments may be written or oral. There will be no formal meeting convened of the project's Advisory Board.

Richard Furuta

Professor, Texas A&M University

Title and Institutional Affiliation

9/22/11

Date

Letter of Commitment to join the Advisory Board for *The Visual Page*

I hereby agree to participate as a member of the Advisory Board for *The Visual Page*, a Start-Up Grant project being proposed to the NEH Office of Digital Humanities by Natalie M. Houston (Associate Professor, Department of English, University of Houston).

If funded, the term of the project and my participation in its Advisory Board will be from May 2012 – May 2013. There will be no remuneration for my participation on this Board.

As a member of the Advisory Board, I will receive quarterly updates on the project's research. I agree to provide comments to the project participants then or at other intervals during the project. Comments may be written or oral. There will be no formal meeting convened of the project's Advisory Board.

Linda K. Hughes
Linda K. Hughes

Addie Levy Professor of Literature, Texas Christian University
Title and Institutional Affiliation

21 September 2011
Date

Letter of Commitment to join the Advisory Board for *The Visual Page*

I hereby agree to participate as a member of the Advisory Board for *The Visual Page*, a Start-Up Grant project being proposed to the NEH Office of Digital Humanities by Natalie M. Houston (Associate Professor, Department of English, University of Houston).

If funded, the term of the project and my participation in its Advisory Board will be from May 2012 – May 2013. There will be no remuneration for my participation on this Board.

As a member of the Advisory Board, I will receive quarterly updates on the project's research. I agree to provide comments to the project participants then or at other intervals during the project. Comments may be written or oral. There will be no formal meeting convened of the project's Advisory Board.

Unmil P. Karadkar
____
NAME

Lecturer, School of Information, The University of Texas at Austin
____
Title and Institutional Affiliation

9/20/2011
____
Date

Letter of Commitment to join the Advisory Board for *The Visual Page*

I hereby agree to participate as a member of the Advisory Board for *The Visual Page*, a Start-Up Grant project being proposed to the NEH Office of Digital Humanities by Natalie M. Houston (Associate Professor, Department of English, University of Houston).

If funded, the term of the project and my participation in its Advisory Board will be from May 2012 – May 2013. There will be no remuneration for my participation on this Board.

As a member of the Advisory Board, I will receive quarterly updates on the project's research. I agree to provide comments to the project participants then or at other intervals during the project. Comments may be written or oral. There will be no formal meeting convened of the project's Advisory Board.

(electronic signature verified by email: L J. Kooistra
Lorraine Janzen Kooistra

Professor of English and Undergraduate Program Director, Ryerson University, Toronto
Title and Institutional Affiliation

September 23, 2011
Date

Letter of Commitment to join the Advisory Board for *The Visual Page*

I hereby agree to participate as a member of the Advisory Board for *The Visual Page*, a Start-Up Grant project being proposed to the NEH Office of Digital Humanities by Natalie M. Houston (Associate Professor, Department of English, University of Houston).

If funded, the term of the project and my participation in its Advisory Board will be from May 2012 – May 2013. There will be no remuneration for my participation on this Board.

As a member of the Advisory Board, I will receive quarterly updates on the project's research. I agree to provide comments to the project participants then or at other intervals during the project. Comments may be written or oral. There will be no formal meeting convened of the project's Advisory Board.

Jerome McGann

John Stewart Bryan University Professor

Title and Institutional Affiliation

21 Sept. 2011

Date

Letter of Commitment to join the Advisory Board for *The Visual Page*

I hereby agree to participate as a member of the Advisory Board for *The Visual Page*, a Start-Up Grant project being proposed to the NEH Office of Digital Humanities by Natalie M. Houston (Associate Professor, Department of English, University of Houston).

If funded, the term of the project and my participation in its Advisory Board will be from May 2012 – May 2013. There will be no remuneration for my participation on this Board.

As a member of the Advisory Board, I will receive quarterly updates on the project's research. I agree to provide comments to the project participants then or at other intervals during the project. Comments may be written or oral. There will be no formal meeting convened of the project's Advisory Board.

Bethany Nowviskie

Director, Digital Research & Scholarship, University of Virginia Library

Title and Institutional Affiliation

21 September 2011

Date

9. Appendices

Appendix A: Selected Visual Features of Printed Poetry

- white space at the top, bottom, inner, outer margins of page
- white space between main text block and running heads or footers
- size and placement of main text block
- size and placement of different levels of titles (section, group, poem, subtitle)
- general page layout (one or multiple columns)
- text density: how crowded or sparse the page seems visually
- white space (leading) between stanzas
- line indentation
- poem separators: white space, ornaments, titles, page breaks (i.e., the first page of a poem often looks different from subsequent pages)
- arrangement, size, placement of running head
- size and placement of page numbers
- size and placement of headnotes or footnotes
- size and features of initial capital letters (i.e., first word of first line may have ornamented or dropped capitalization)
- typeface size(s) on the page
- mixed typefaces or typeface attributes on the page
- size and placement of illustrations or ornaments

Appendix B: Technical Overview

The *Visual Page* application will work in three main stages. In the first stage, page layout analysis algorithms are used to identify significant visual features of the page. See Appendix A for a list of potential visual features. In the second stage, machine learning and pattern analysis algorithms are applied in order to detect interesting relationships and discover meaningful patterns. The third stage consists of presenting these relationships to scholars for analysis and to enable them to interactively define and revise the questions they pose to their data.

Feature Identification

The first step, feature identification, refers to the process of defining which visual features are important for scholarly analysis and building algorithms that are capable of identifying these features within a document image. The components of the application that do this are called feature identification algorithms or feature identifiers for short. For example, simple feature identifiers could be built to recognize the number of lines on a page, line height (as a percentage of the page size), and the spacing between lines (as a percentage of the average line height). More advanced features identifiers might determine which typefaces were used on the page, the use of bold or italics, or how regular the left or right margin is.

To achieve this, we will rely heavily on prior work in document layout analysis (Mao et al. 2003; Shafait et al. 2008). Layout analysis is a mature field with a number of successful, well-known algorithms (e.g., Nagy et al. 1992; O'Gorman 1993; Kise et al. 1998) as well as more advanced approaches designed to solve specific problems (Ramel 2007; Smith 2009). This prior work will allow us to implement image analysis algorithms capable of identifying lines on the page, blocks of text such as stanzas, titles, running headers, marginal notes, white-space regions and more. The technological innovations of our project come from the way these algorithms are used. Existing work focuses on the comparatively concrete (and financially rewarding) task of using optical character recognition to extract the linguistic meaning of the page. We will use this technology to provide scholars with tools to help analyze the culturally conditioned ways that documents convey meaning through their visual structure.

Rather than defining a final set of features to be used in analyzing a page, we will design our application so that someone with suitable programming ability can create a new feature identifier. This will allow the application to incorporate technological advances in the field of image analysis and to support scholarly innovation as humanists identify new visual features that they would like to investigate. Thus, we envision the final *Visual Page* application not merely as a stand-alone tool, but as a software development environment that enables scholars to work with technologists in order to continuously extend the system to ask new questions.

These features, once recognized from a page image, are represented in a machine readable form as a feature vector. Just as we can identify a point in a three dimensional space with its x, y and z coordinates (latitude, longitude and altitude for a geographical example), we can create a model of the visual space in which documents exist by treating each recognized feature as a single dimension. Instead of three dimensions, this creates a high-dimensional vector space in which each visual feature that we chose to analyze is a dimension and the documents form individual points in this space.

This is the same approach used by modern search engines. Search engines work by representing each word as a dimension in a space with hundreds of thousands of dimensions. Documents are located in this space by counting how frequently a particular word appears in the document. A search then finds documents that are closest to a specific query in this space. Document clustering works by grouping documents that are near each other in this space.

As a model of documents, this approach emphasizes certain features of the document's meaning while simultaneously ignoring others (McCarty 2008). In *The Visual Page*, we are privileging visual indicators of meaning over verbal indicators. By providing a formal mechanism to represent the visual features that scholars find important, they will be able to ask questions about the visual appearance of a page systematically over large document collections in much the same way that search engines and text analysis tools allow them to ask systematic questions about a document's linguistic content.

Analyzing Relationships

Once documents have been modeled using feature vectors, we can apply many different pattern classification algorithms to help discover potential relationships, to reveal patterns that might be difficult to detect through manual inspection, or to provide systematic confirmation (or rejection) for intuitions developed by looking at a few exemplary works. For the prototype system, we will evaluate three broad pattern classification techniques in the context of specific research questions:

Clustering. Clustering is a an unsupervised approach to grouping documents based on the distances between their feature vectors.[1] This pattern classification technique allows the computer to find groupings of visually similar texts based on the features that have been selected by scholars as important dimensions of visual meaning. Once documents have been grouped into clusters, we can ask questions like

- What do average or nearly average documents in this group look like?
- How much diversity is there in marginal width or text density within this group?
- Are there significant factors (such as a common publisher, time period, geographic location or genre) that explain this visual similarity?
- How different is this group from another?

Clustering does not need to be performed on the entire collection or the entire set of visual features. Instead, a scholar could select a sub-set of features to analyze in order to ask questions specifically about the use of white space or alternatively the use of typographical features such as the typeface used, line height, and line spacing. Another option would be to identify clusters within a specific time span (the 1860s, for example) and then track those clusters as they change over time using a tool like Kalman filters.

These algorithms and the resulting clusters can support sophisticated measures of similarity and difference. For example, when comparing the similarity of two different groups, we want to use not just the computed average for the cluster but to combine measures for "within-class scatter"and "between class scatter" That is, two groups in which all the documents within a group are very similar but whose centers are relatively close may be more distinct than two groups whose centers are far apart, but where there is a lot of diversity between documents within a group.

[1] We plan to develop tools that use expectation maximization (EM) to fit Gaussian models to the documents in our collection. We will perform this clustering in both in the base vector space and in a reduced dimensional space computed using Principal Components Analysis (PCA).

Classification. Classification algorithms are a supervised approach to assigning a meaningful class label to a group of documents.[2] For example, someone might want to find all pages with sonnets based on the visual features of the page image. These algorithms are said to be supervised because they need to be trained with correct examples before they can work. To recognize sonnets, you would begin by providing 100 example pages that contained sonnets and 100 pages that contained a selection of other material.

Classification algorithms can be used to ask questions in the form of "find me all documents that look like this but not like that." For example, these could be used to distinguish between documents that look like those published by Macmillan versus those published by Bell and Daldy. A more sophisticated application of classification would also allow a researcher to gather documents that she found to be visually similar (these will be assigned the class "desired" and to retrieve documents that looked more like these desired documents than the average document. This initial set of documents will include some documents the researcher is not interested in, so she could then refine the classification algorithm by identifying "undesirable" documents that would be used to retrain the classifier and produce better results.

As a more concrete example, the researcher might take Tennyson's *In Memoriam* as a starting point to discover possible patterns of imitation or to identify prior examples of visually similar documents. Upon retrieving the 100 most similar documents, the researcher would inspect the results to identify works that were, in her judgment, relevantly similar and relevantly dissimilar. The system could then train a classifier to distinguish between these two sets and return other similar documents. It would also be able to provide statistics about how accurate the classifier is through a technique called cross-validation.[3]

Feature Selection. Working with high-dimensional spaces is problematic for a number of reasons. One of these reasons is that some features are likely to be highly relevant to a particular set of questions, while others are just noise. There are statistical models that are designed to reduce noise in the data and help smooth out differences between different dimensions in the data set, but these typically result in a new set of dimensions that, while useful for computational algorithms like clustering and classification, do not translate into meaningful visual properties of the image.

Feature selection techniques help address this problem by evaluating the performance of a classifier on different sub-sets of features in order to determine which features are the most effective for a particular classification problem. Since it is usually not possible to evaluate all combinations of features, these algorithms use a heuristic approach to estimate which features are most important.[4] We propose to use feature selection to identify the relevant visual properties that distinguish different groups of documents. For example, we might use feature selection to indicate which visual features were most important in differentiating *In Memoriam* from other books published in 1850.

[2] We plan to use Linear Discriminant Analysis and Maximum Likelihood classifiers in our prototype system. We may use other classifiers depending on the need to answer specific research questions and available time.

[3] This involves using some of the training data to train the classifier and reserving a portion of the training data to evaluate its accuracy. This technique can be used to select the best classifier for a particular set of data or to provide insight on the expected accuracy of the classifier.

[4] An exhaustive evaluation to find the 10 most important features out of 20 involves 184,756 feature subsets. Evaluating each sub-set requires training a classifier for that sub-set and testing its performance.

Data Visualization

Tools to extract document features and identify their relationships are only useful if there are ways to present those relationships to scholars so that they can systematically evaluate hypotheses formed on the basis of manual inspection and gain additional insights that are difficult to reach by examining documents individually. Consequently, techniques and tools to visualize the results of the computational analyses will drive the success of this project in facilitating scholarly inquiry. Significant work has been done in the field of information visualization in general (Card, et al. 1999) and for visually significant cultural heritage data in particular (i.e., *ImagePlot*). We expect that we can easily export our data in formats that can be ingested by many of these existing systems. Development of visualization tools will be a major component of future work (beyond the Start-Up Grant period) as we move from our proposed prototype to a more mature application.

Within the scope of our prototype work, we will focus on the use of tiled thumbnails, montage, and filmstrip displays as tools to help scholars see and understand visual relationships between groups of images. Preliminary experiments with scientific images (Karadkar, et al, 2006) have shown that these techniques help provide contextual information between images through spatial (thumbnail display) or temporal (montage) juxtapositions, or a combination of both (scrolling filmstrip). We plan to explore the use of interactive visualizations in depth when implementing the final system both by applying existing tools and developing custom interfaces as needed. Time permitting, we will begin this work during the Start-Up phase.

Summary

The *Visual Page* will allow researchers to pose new questions that engage both the linguistic and graphic signs that jointly convey meaning in printed texts. Research in document image analysis has been incredibly successful in extracting linguistic content while treating the graphical elements of these documents as technical difficulties to be overcome. Consequently, the visual components of a page remain invisible to computers. Our work will enable computers to see many of the graphic features that humanities scholars use to evaluate documents and to provide computational support for understanding how those features convey meaning.

We envision a system that will enable people to select the visual properties and other metadata fields to be analyzed, use pattern analysis and other machine learning techniques to pose questions, and incrementally revise and update their queries as they gain a deeper understanding of both the documents they are studying and the research questions they are pursuing. This Level II Start-Up Grant will lay the foundation for this work as we design initial feature identification and pattern analysis algorithms. A key advantage of the basic strategy that we propose to adopt, feature extraction followed by pattern analysis, is that it can be easily extended to different types of documents and allow researchers to pose new questions. As new visual features need to be represented, this can be done by implementing a new feature identification algorithm capable of recognizing this feature and extending the existing feature vectors for the document images. Beyond the machine learning techniques that we may develop ourselves, the feature vectors for page images can be exported to formats like the Attribute-Relation File Format (ARFF)[5] used by Weka (Hall, et al. 2009). By decoupling the image analysis stage from the pattern analysis stage, our work will allow researchers to mix and match the tools that best suit their needs and also minimize the need to duplicate functionality already provided by existing systems.

[5] http://weka.wikispaces.com/ARFF

Appendix C: References

Anderson, Benedict. *Imagined Communities: reflections on the origin and spread of nationalism*. London: Verso, 1983.

Audenaert, Neal, Unmil Karadkar, Enrique Mallen, Richard Furuta, and Sarah Tonner. "Viewing Texts: An Art-Centered Representation of Picasso's Writings." Digital Humanities 2007, Urbana-Champain, IL (June 4-7, 2007).

Audenaert, Neal and Richard Furuta. "Annotated Facsimile Editions: Defining Macro-level Structure for Image-Based Electronic Editions." *Literary and Linguistic Computing* 24.2 (2009): 143-151.

Audenaert, Neal, George Lucchese, and Richard Furuta. "CritSpace: A workspace for critical engagement within cultural heritage digital libraries." Proceedings of ECDL 2010 LNCS 6273. Berlin: Springer, 2010. 307-314.

Audenaert, Neal and Richard Furuta. "What humanists want: how scholars use primary source documents." Proceedings of the 10th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2010, Gold Coast, Australia (June 21-25, 2010): 283-292.

Benton, Megan. *Beauty and the Book: Fine Editions and Cultural Distinction in America*. New Haven: Yale UP, 2000.

Bishop, Christopher M. *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

Calè, Luisa and Patrizia di Bello, eds. *Illustrations, Optics and Objects in Nineteenth-Century Literary and Visual Cultures*. Basingstoke: Palgrave Macmillan, 2010.

Card, Stuart K., Jack D. Mackinlay and Ben Shneiderman, eds. *Readings in Information Visualization: Using Vision to Think*. San Francisco: Morgan Kaufmann, 1999.

Chartier, Roger. *The Cultural Uses of Print in Early Modern France*. Trans. Lydia Cochrane. Princeton: Princeton UP 1987.

Cohen, Dan. "Searching for the Victorians." http://www.dancohen.org/2010/10/04/searching-for-the-victorians

Darnton, Robert. "What Is the History of Books?" *Daedalus* 111.3 (Summer 1982): 65-83.

Dowling, Linda. "Letterpress and Picture in the Literary Periodicals of the 1890s." *The Yearbook of English Studies* 16 (1986): 117-131.

Drucker, Johanna. "Not Sound." *The Sound of Poetry: The Poetry of Sound*. Ed. Marjorie Perloff and Craig Dworkin. Chicago: U of Chicago P, 2009. 237-48.

-----. *SpecLab: Digital Aesthetics and Projects in Speculative Computing*. Chicago: U of Chicago P, 2009.

Eisenstein, Elizabeth. *The Printing Press as an Agent of Change: Communications and Cultural Transformations in Early Modern Europe*. Cambridge: Cambridge UP, 1979.

Enenkel, Karl and Wolfgang Neuber, eds. *Cognition and the Book: Typologies of Formal Organisation of Knowledge in the Printed Book of the Early Modern Period*. Leiden: Brill, 2005.

Erickson, Lee. *The Economy of Literary Form: English Literature and the Industrialization of Publishing, 1800-1850*. Baltimore: Johns Hopkins UP, 1996.

Frankel, Nicholas. *Oscar Wilde's Decorated Books*. Ann Arbor: U of Michigan P, 2000.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten. "The WEKA Data Mining Software: An Update" *SIGKDD Explorations* 11.1 (November 2009): 10-18.

Hayles, N. Katherine. *My Mother was a Computer: Digital Subjects and Literary Texts*. Chicago: U of Chicago P, 2005.

Helsinger, Elizabeth. *Poetry and the Pre-Raphaelite Arts: Dante Gabriel Rossetti and William Morris*. New Haven: Yale UP, 2008.

Hilliard, Christopher. *To Exercise our Talents: the Democratization of Writing in Britain*. Cambridge: Harvard UP, 2006.

Hoagwood, Terence and Kathryn Ledbetter. *Colour'd Shadows: Contexts in Publishing, Printing, and Reading Nineteenth-Century British Women Writers*. New York: Palgrave Macmillan, 2005.

Houston, Natalie M. "Newspaper Poems: Material Texts in the Public Sphere." *Victorian Studies* 50.2 (Winter 2008): 233-242.

-----. "Valuable by Design: Material Features and Cultural Value in Nineteenth-Century Sonnet Anthologies." *Victorian Poetry* 37 (Summer 1999): 243-272.

Johns, Adrian. *The Nature of the Book: Print and Knowledge in the Making*. Chicago: U of Chicago P, 1998.

Karadkar, Unmil, Marlo Nordt, Richard Furuta, Cody Lee, and Christopher Quick. "An exploration of space-time constraints on contextual information in image-based testing interfaces." Proceedings of ECDL 2006 LNCS 4172. Berlin: Springer, 2006. 391-402.

Kise, Koichi, Akinori Sato, and Motoi Iwata, "Segmentation of page images using the area Voronoi diagram." *Computer Vision and Image Understanding* 70.3 (June 1998): 370-382.

Kooistra, Lorraine Janzen. *Poetry, Pictures, and Popular Publishing: The Illustrated Gift Book and Victorian Visual Culture, 1855-1875*. Athens: Ohio UP, 2011.

Levenston, Edward A. *The Stuff of Literature: Physical Aspects of Texts and Their Relation to Literary Meaning*. Albany: State U of New York P, 1992.

McGann, Jerome J. *Radiant Textuality: Literature after the World Wide Web*. New York: Palgrave Macmillan, 2001.

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

McGill, Meredith L. *American Literature and the Culture of Reprinting, 1834-1853*. Philadelphia: U of Pennsylvania P, 2003.

McKenzie, Donald Francis. *Bibliography and the Sociology of Texts*. Cambridge: Cambridge UP, 1999.

Mandell, Laura. "What is the Matter? What Literary Theory Neither Hears nor Sees." *New Literary History* 38.4 (Autumn 2007): 755-76.

Manguel, Alberto. *A History of Reading*. New York: Viking, 1996.

Mao, Song, Azriel Rosenfeld and Tapas Kanungo. "Document structure analysis algorithms: a literature survery." Proceedings of SPIE Electronic Imaging 5010 (January 2003): 197-207.

Maruca, Lisa. *The Work of Print: Authorship and the English Text Trades, 1660-1760*. Seattle: U of Washington P, 2007.

Mitchell, Tom M. *Machine Learning*. Boston: McGraw-Hill, 1997.

Nagy, George, Sharad Seth, and Mahesh Viswanathan, "A prototype document image analysis system for technical journals." *Computer* 25.7 (July 1992): 10-22.

Nagy, George. "Twenty years of document image analysis in PAMI" IEEE Transactions on Pattern Analysis and Machine Intelligence 22.1 (January 2000): 38-62.

O'Gorman, Lawrence. "The document spectrum for page layout analysis." *IEEE Trans. Pattern Analysis and Machine Intelligence* 15.11 (November 1993): 1162-1173.

Ong, Walter J. *Orality and literacy: the technologizing of the word*. London: Methuen, 1982.

Ramel, J. Y., S. Leriche, M. L. Demonet and S. Busson. "User-driven page layout analysis of historical printed books." *Intl. Journal on Document Analysis and Recognition* 9.2-4. (2007): 243-261.

Raven, James, Helen Small, and Naomi Tadmor, eds. *The Practice and Representation of Reading in England*. Cambridge: Cambridge UP, 1996.

Saenger, Paul Henry. *Space Between Words: the Origins of Silent Reading*. Stanford: Stanford UP, 1997.

St. Clair, William. *The Reading Nation in the Romantic Period*. Cambridge: Cambridge UP, 2004.

Shafait, Faisal, Daniel Keysers and Thomas M. Breuel. "Performance evaluation and benchmarking of six-page segmentation algorithms." *IEEE Trans. Pattern Analysis and Machine Intelligence* 30.6 (June 2008): 941-954

Sher, Richard B. *The Enlightenment & the Book: Scottish Authors and their Publishers in Eighteenth-Century Britain, Ireland, and America*. Chicago: U of Chicago P, 2006.

Smith, Margaret. *The Title-Page: its Early Development, 1460-1510*. London: British Library, 2000.

Smith, Ray. "Hybrid Page Layout Analysis via Tab-Stop Detection." Proceedings of the 10th International Conference on Document Analysis and Recognition ICDAR 2009, Barcelona, Spain (2009): 241-245

Urbina, Eduardo, Richard Furuta, Steven Escar Smith, Neal Audenaert, Jie Deng, and Carlos Monroy. "Visual Knowledge: Textual Iconography of the Quixote, a Hypertextual Archive." *Literary and Linguistic Computing* 21.2 (June 2006): 247-258.

Viscomi, Joseph. "Digital Facsimiles: Reading the William Blake Archive." *Computers and the Humanities* 36.1 (February 2002): 27-48.