



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: TOME: Interactive Topic Model and METadata Visualization

Institution: Georgia Tech Research Corporation

Project Directors: Lauren Frederica Klein

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

# NEH Application Cover Sheet

## Digital Humanities Start-up Grants

### PROJECT DIRECTOR

---

Dr. Lauren Frederica Klein  
Assistant Professor  
Skiles 359, School of Literature, Media and  
Culture  
Georgia Institute of Technology  
Atlanta, GA 303320420  
UNITED STATES

**E-mail:** lauren.klein@lmc.gatech.edu

**Phone(W):** 917-887-2379

**Phone(H):**

**Fax:**

**Field of Expertise:** Languages - English

### INSTITUTION

---

Georgia Tech Research Corporation  
Atlanta, GA UNITED STATES

### APPLICATION INFORMATION

---

**Title:** *TOME: Interactive TOPic Model and METadata Visualization*

**Grant Period:** From 5/2013 to 4/2014

**Field of Project:** Humanities

**Description of Project:** As archives are being digitized at an increasing rate, scholars will require new tools to make sense of this expanding amount of material. We propose to build TOME, a tool to support the interactive exploration and visualization of text-based archives. Drawing upon the technique of topic modeling—a computational method for identifying themes that recur across a collection—TOME will visualize the topics that characterize each archive, as well as the relationships between specific topics and related metadata, such as publication date. An archive of 19th century antislavery newspapers, characterized by diverse authors and shifting political alliances, will serve as our initial dataset; it promises to motivate new methods for visualizing topic models and extending their impact. In turn, by applying our new methods to these texts, we will illuminate how issues of gender and racial identity affect the development of political ideology in the nineteenth century, and into the present day.

### BUDGET

---

<b>Outright Request</b>	\$59,999.00	<b>Cost Sharing</b>	
<b>Matching Request</b>		<b>Total Budget</b>	\$59,999.00
<b>Total NEH</b>	\$59,999.00		

### GRANT ADMINISTRATOR

---

Mrs. Stacey Oliver-Gooden  
Contracting Officer  
Office of Sponsored Programs  
505 Tenth Street  
Atlanta, GA 30332-0420  
UNITED STATES

**E-mail:** sogooden@gatech.edu

**Phone(W):** 404-894-6930

**Fax:** 404-894-5945

**TOME:**  
***Interactive TOPic Model and METadata Visualization***

**A Level II Start-Up Project proposed by the  
Digital Humanities Lab at the Georgia Institute of Technology**

<b>1. Table of Contents</b>	
<b>2. List of Participants</b>	<b>2</b>
<b>3. Abstract</b>	<b>3</b>
Statement of Innovation	3
Statement of Humanities Significance	3
<b>4. Narrative</b>	<b>4</b>
Introduction: Enhancing the Humanities through Innovation	4
Environmental Scan	4
Project Description	6
<i>The TOME Interface</i>	6
<i>Research Questions</i>	6
History and Duration of Project	8
Work Plan	8
Staff	9
Final Product and Dissemination	9
<b>5. Project Budget</b>	<b>10</b>
Budget Narrative	10
<b>6. Biographies</b>	<b>12</b>
Jacob Eisenstein	12
Lauren Klein	12
Janet Murray	12
John Stasko	12
<b>7. Data Management Plan</b>	<b>13</b>
Data to be Generated	13
Period of Data Retention	13
Data Formats and Dissemination	13
Data Management and Maintenance	13
<b>8. Letters of Support</b>	<b>14</b>
Amy Earhart	15
David Mimno	17
<b>9. Appendices</b>	<b>18</b>
Appendix A: Topic Modeling Background	18
Appendix B: Modeling Influence	18
Appendix C: Text Visualization	19
Appendix D: Public Dataset	20
Appendix E: Works Cited	21

## 2. List of Participants

Eisenstein, Jacob	Georgia Institute of Technology
Klein, Lauren	Georgia Institute of Technology
Murray, Janet	Georgia Institute of Technology
Stasko, John	Georgia Institute of Technology

### 3. Abstract

As archives are being digitized at an increasing rate, scholars will require new tools to make sense of this expanding amount of material. We propose to build TOME, a tool to support the interactive exploration and visualization of text-based archives. Drawing upon the technique of topic modeling—a computational method for identifying themes that recur across a collection—TOME will visualize the topics that characterize each archive, as well as the relationships between specific topics and related metadata, such as publication date. An archive of 19<sup>th</sup>-century antislavery newspapers, characterized by diverse authors and shifting political alliances, will serve as our initial dataset; it promises to motivate new methods for visualizing topic models and extending their impact. In turn, by applying our new methods to these texts, we will illuminate how issues of gender and racial identity affect the development of political ideology in the nineteenth century, and into the present day.

#### Statement of Innovation

Topic modeling research continues to expand, yet there has been little attention to how topic models can be used to enhance humanistic inquiry. Driven by fundamental humanities questions about the evolution and circulation of ideas, we will develop new methods for combining topic models with metadata, and new modes of interactive visualization for exploring the results. These will illuminate new connections across social networks and over time, while remaining accessible to non-technical users.

#### Statement of Humanities Significance

The antislavery newspapers of the 19th-century United States are significant not only because they function as the primary record of the sociopolitical machinations that culminated in slavery's abolition, but also because they are among the earliest examples of women writing alongside men for a general audience. Our project will reveal new ways in which ideas passed between social and political coalitions, with an emphasis on how these ideas were framed differently by male and female writers.

## 4. Narrative

### Introduction: Enhancing the Humanities through Innovation

Examining the origin, evolution, and circulation of ideas is among the most fundamental tasks of humanities scholarship. The digitization of archival collections has greatly increased the amount of primary material for such research, but scholars have struggled to sift through these ever-expanding online archives. We believe that computational analysis can help to bridge the gap between digitized text and scholarly interpretation, much as it has helped biologists to more efficiently process the voluminous data made available by automated gene-sequencing techniques. To this end, we propose to develop a web-based tool for the visual exploration of the themes that recur across an archive, based on the text-analysis technique of *topic modeling*. In so doing, we will enable humanities scholars to trace the evolution and circulation of these themes across social networks and over time.

Our initial focus is on a set of nineteenth-century abolitionist newspapers, penned between 1830 and 1865, in which United States antislavery advocates mounted moral, social, and political arguments in favor of a general emancipation. These texts, in the public domain and available in digitized, machine-readable form through the ProQuest “Black Abolitionist Papers” database (see “Data Management Plan”), will serve as our sample set for constructing and refining our computational approach. With a prototype of TOME, our web tool, to be completed within the grant-funded year, we hope to illuminate the pathways of ideas and influence across these important newspapers. Indeed, these pathways were fraught—for as many ideas were shared by this diverse group of writers, which included men and women, African Americans and whites, and Northerners and Southerners, they often disagreed as to how best to achieve their common goal. While historians and cultural critics have examined the reverberations of dramatic events, such as when Lydia Maria Child resigned her position as editor of the *National Anti-Slavery Standard* after refusing to serve up the “hyena soup with brimstone seasoning” that her more radical friends desired, we hope to identify additional events and ideas, and other aspects of ideological influence, that might have traveled through these papers undetected, and that might complicate our current understanding of that momentous time (Qtd. in Karcher 1997).

More specifically, TOME will allow scholars to address the following questions: (1) **What** are the main themes in this document collection? (2) **Who** are the authors and subjects most closely associated with each theme? (3) **When** was each theme most prominent? (4) **How** did each theme spread from its source to the wider community? While our humanities research interests lie in this set of abolitionist newspapers, and in particular, in the relation of ideological influence to gender, we believe TOME will present wide-ranging uses for humanities scholars. Indeed, the second phase of our project will involve the creation of a web-based tool to support non-technical users throughout the entire lifecycle of topic-based analysis for any set of digitized documents. As such, TOME will be the first tool to provide a visual interface for exploring the relationships among textual content and related metadata. In the pages that follow, we describe these innovations, as well as our approach to answering the above research questions in more detail.

### Environmental Scan

In recent years, scholars from across the humanities have begun to investigate the applicability of *topic modeling* – that is, a statistical technique for automatically identifying the themes that recur across a document collection – to humanities scholarship. This has yielded new insights about literary diction (Underwood 2012) and poetic convention (Rhody 2012), as well as thematic trends in novels (Jockers 2013) and in historical accounts (Blevins 2010). While

topic modeling algorithms are implemented in existing software libraries such as MALLET (McCallum 2002) – indeed, this is the software that Rhody, Jockers, and Blevins each employ – such libraries provide little support for exploring, or even understanding the themes they extract; nor do these libraries allow users to make connections to important metadata, such as author or publication date. For the humanities scholar, this makes topic-based research incredibly technically challenging, not to mention time consuming. TOME will help non-technical users to visualize and explore the output of topic modeling algorithms through an innovative interactive interface, which we discuss in more detail in our “Project Description.”

At present, a few websites provide interactive interfaces for exploring the output of topic models, although each site is designed around a specific archive. “Mapping Texts” displays lists of topics associated with a series of historical time-periods (Yang et al., 2011). However, there is little integration of the topics with the document metadata, or with the documents themselves, thereby limiting the utility of the topics to facilitate additional research. “The Open Encyclopedia of Classical Sites” (Mimno 2012) displays the topics associated with a set of locations referenced in data from Google Books. It also provides visualizations of the metadata associated with each of the topics: spatial information is shown using the Google Maps API, and temporal strength is shown with a graph. But this site, like “Mapping Texts,” displays a limited list of topics, each represented by a series of words. It is not clear how this interface would scale to models involving more topics, since current models result in hundreds or even thousands of topics (see Mimno et al. 2012). To augment these approaches, we propose to develop an interactive visualization that will allow scholars, through an iterative process, to refine the topics and texts most relevant to their research.

Considering computational text-analysis more broadly, there are several significant efforts that complement our proposed work. The MONK and WordSeer projects each use automated syntactic analysis to identify named entities (MONK), part-of-speech sequences (MONK and WordSeer), and grammatical dependency relations (WordSeer). Like the topic modeling techniques we propose to employ, automated syntactic analysis can also be seen as a method that improves upon more basic word-counting models by adding context: syntactic analysis accounts for context at the level of individual sentences, while topic analysis accounts for context across the entire document. Another key difference from syntactic analysis is that topic models do not require labeled training data, thus increasing their applicability to diverse writing styles and genres. In any case, the success of projects like MONK and WordSeer attests to the desire, on the part of humanities scholars, for better text-exploration tools. We intend to use these projects, their documentation, and the methods by which they have introduced themselves to the digital humanities community, as models for our project as we prepare for its public release.

As far as the humanistic aspect of our project, numerous studies have explored the complex relationships among the various (and often competing) antislavery newspapers. As early as David Brion Davis’s 1979 anthology of primary sources, *Antebellum American Culture*, and even before, these newspapers have been recognized for their profound influence, both individually and collectively, on U.S. politics. In response to the waning of second-wave feminism, and to the contemporary political climate, recent scholarship has sought to renew the focus on the role of women writers in advocating for slavery’s abolition. Christine Stansell’s *The Feminist Promise* (2010) and Faye Dudden’s *Fighting Chance* (2011) are among the works that devote significant attention to the antislavery newspapers in which these women wrote. Carla Peterson’s award-winning *Black Gotham* (2011) builds its cultural history in large part from the very newspapers we seek to explore. By combining topic analysis with metadata, we hope to lend additional insight into how these women affected the politics of their time, and consequently, of today.

## Project Description

We now present the design principles behind the TOME interface and provide an overview of its basic functionality. We then describe the specific methods of analysis and forms of visualization we will employ in order to address our four research questions.

### ***The TOME Interface***

Traditional keyword-based interfaces return a set of documents whose content matches a user query. The strength of such interfaces is that they have a well-understood workflow, allowing the researcher to quickly locate individual documents of interest. However, search-based interfaces require the query terms to be known in advance, making exploratory research more difficult. Spatial text visualizations (see Appendix C) have a complementary set of advantages and disadvantages. While they are able to show the large-scale thematic differences among documents in a collection, they must compress the entire thematic space of the collection into a single view. In addition, these visualizations tend to be static, and do not permit viewers to focus on the thematic contrasts most relevant for their research.

TOME will bridge the gap between these two techniques. The interaction begins with a keyword query—say, for “emancipation”—but rather than return a list of documents, as in a traditional search interface, TOME will display the documents in an interactive spatial layout, organized by the topics that best match the initial query results. The user can then navigate the thematic landscape, identifying the topics relevant to her research along with their metadata properties. For example, the documents relating to “emancipation” might focus on three primary themes: political measures, moral arguments, and supporting institutions. TOME will allow the scholar to rapidly identify the documents that focus on each of these themes, and to refine the working set of documents – for example, by removing all documents whose main emphasis is “moral arguments.”

This visual, topic-based method of exploration allows the scholar to quickly build and refine document sets based on their thematic characteristics, rather than assembling large and possibly error-prone lists of keywords. It can also help to reveal additional search terms that might not have been considered at the outset. In the case of public institutions developed in the wake of emancipation, for instance, the scholar might think to search for “school” and “hospital,” but might not think to search for “prison.” As a result, she would not see the documents that describe the institution that arguably carries the most profound implications for the present day. We thus see the need for an iterative process, in which the user identifies an initial set of search terms, and then employs the visualization in order to explore the most relevant topics and to discover thematically-relevant documents that might have been missed in the initial query.

### ***Research Questions***

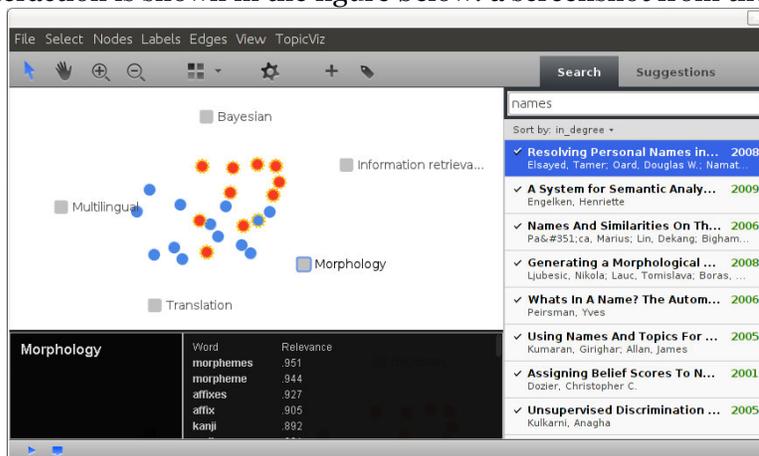
The TOME interface will allow scholars to address four specific questions relating to the nature of the topics contained within the archive, and to the links between topics and various metadata. We believe these questions can serve as the starting point for deeper questions about content, authorship, circulation, and influence. In the discussion that follows, we describe how TOME will address each question, and give examples of how TOME might serve as the basis for more nuanced cultural analysis.

### **WHAT are the main themes in the document collection?**

The starting point of any theme-based research is to determine the topics in the collection, and the words and documents associated with those topics. TOME will reveal these associations through a *dust-and-magnet visualization* (Yi et al, 2005). In this mode of visualization, each topic acts as a “magnet,” exerting force on the other elements of the

visualization: either words or documents. The stronger the relationship between the topic and the document (or word), the greater the force that the topic magnet exerts. By re-arranging the topic magnets, the user can create layouts that foreground specific comparisons or contrasts between topics. By adding or removing topics or documents, the user can drill down to more focused sections of the archive.

An example of this type of interaction is shown in the figure below: a screenshot from the TopicViz system for exploring scientific research literature (Eisenstein et al. 2012). In this example, the user has already entered a search query and retrieved a set of documents; she is presented with a spatial visualization of the relationship between her search terms and the five most relevant topics. In TopicViz, topics are shown by lists of relevant words (see the bottom center panel), but our research on TOME will investigate visual presentations that use the same form of dust-and-magnet visualization to display the words that are associated with each topic. This will allow the scholar to focus her research on the words most relevant to her research.



### WHO are the authors most associated with each theme?

Analyses of topic models of historical and literary texts have tended to focus on the words that characterize each topic. However, the topic model also computes the topical composition of each document. In the example of “emancipation,” given above, the topic model may determine a particular document to be composed of 10% political measures, 30% moral arguments, and 60% supporting institutions. A powerful application of topic models combined with metadata would be to aggregate the topics of all documents composed by the same author, or set of authors, thus allowing scholars to compare and contrast each author's thematic interests.

TOME will visualize these comparisons in two distinct ways. For a small number of authors, the relevant document nodes in the visualization can be color coded, as in the image above. For a larger numbers of authors, we will invert the layout, using the authors as “magnets” (nodes whose position is determined by the user) and the topics as dust (nodes whose position is determined by the forces exerted by the magnets). A topic that is more closely associated with a particular author will be more strongly attracted to that author's magnet; this will be reflected in the spatial position of the topic's node.

These visualizations will be especially illuminating for our test dataset of abolitionist newspapers, since it is often assumed that women writers tended to focus on domestic themes, such as children, education, or the home, in order to advance their antislavery arguments, while male authors focused more explicitly on political arguments to abolish slavery. Lydia Maria Child, the indignant editor quoted above, once claimed to attract women readers by adorning her “anti-slavery principles” with the “garland of imagination and taste” (Qtd. in Karcher, 1997). But were there women writers who defied this pattern? Or women whose more subtly insinuated ideas were later taken up by men? The visualizations described above will allow us to begin to determine the answer to these questions.

### WHEN was each topic most prevalent?

Identifying the major shifts in favor of (or against) a particular social or political issue has long been a topic of interest for humanities scholars—for example, in the nineteenth century, determining when public support for immediate emancipation eclipsed support for various schemes to expatriate black Americans to Africa. If publication-date metadata is available for the documents in an archive—which it is in the case of newspaper articles—then TOME can measure the popularity of a topic at any time period by averaging the topical composition of each document penned during the period. A visualization of the change in topical composition over time can offer an intuitive sense of how the interests of an author, publication, community, or regional area evolved. Such a visualization can take the form of a simple line graph; alternatively, within the inverted dust-and-magnet layout described above, we can use magnets to represent time periods, and represent the topics as dust, whose position is determined by the prevalence of the topic during each relevant time period. To make more precise statements about the changing popularity of a topic, we can formulate statistical hypothesis tests of how topic strength changes over time.

### **HOW does a topic spread?**

Which individuals, newspapers, or regions were most responsible for the spread of a topic? Although humanities scholars tend to attribute the spread of ideas to a range of factors – influential publications, latent social forces, catalyzing events, etc. – influence in its most basic sense is a form of causality. We propose to determine influence by first identifying which topics originate with a single author, newspaper, or region, and then by determining which other authors—or which other sources—later took up the same topics. Our initial dataset comprises a closely-knit group of writers who actively responded to each others’ work – as when Maria Weston Chapman, one of the more radical voices of the American Antislavery Society, accused a colleague of “substituting ‘flapdoodle’ for the ‘roast Beef’” the paper required (Qtd. in Karcher 1997). For this reason, our initial dataset offers a compelling test case for this model of influence. Not only might this research yield insights into the ways in which women asserted authority and influence over their male peers, it also goes well beyond the capabilities of traditional topic models (see Appendix B for more discussion). As such, it demonstrates the powerful potential of direct collaborative research between scholars from the humanities and computer science.

### **History and Duration of the Project**

This project has emerged from a collaboration that applied topic models to the complete *Papers of Thomas Jefferson*, a subject of Dr. Klein’s research. A paper documenting this project will be published in the proceedings of *Research Foundations for Understanding Books and Reading in the Digital Age: E/Merging Reading, Writing, and Research Practices* in early 2013.

Dr. Eisenstein has performed initial research on the project’s underlying computational methods, including the development of a topic model visualization system for scientific research papers, TopicViz, mentioned above. While TopicViz can serve as a prototype for TOME, significant original research will be required to develop the features proposed in the previous section.

Dr. Klein has identified the set of newspapers that will be included in the initial dataset (see “Data Management Plan”). In addition, she has conducted substantial research on several of these newspapers as part of her current book project, an examination of the interrelation of food and politics—including abolitionism—in the early republic.

### **Work Plan**

With funding to support a graduate student for one year, we can implement visualizations for the “what,” “who,” and “when” of topic models; we will also undertake the

more challenging “how” question, as described above. The startup phase will focus on developing visualizations for the output of existing topic model toolkits. Follow-up funding would enable us to build a more comprehensive tool that would support humanities researchers throughout the lifecycle of a topic-based investigation, including the preprocessing steps required for topic modeling, as well as providing a topic modeling implementation that is easy to use and install.

As the project will continue after the period of the grant, we intend to seek funding from the NEH Digital Humanities Implementation Grant program. We may also consider funding sources outside of the humanities, as we intend to adapt our tool to be used with social media data. These contemporary applications may be funded by Department of Defense programs such as Minerva.

### Staff

**Jacob Eisenstein (Project Co-Director).** Assistant Professor, School of Interactive Computing, Georgia Tech. Dr. Eisenstein is interested in how latent variable models of text can be used to support humanistic insights. He will conduct weekly meetings with his graduate students to guide their implementation of the prototype.

**Lauren Klein (Project Director).** Assistant Professor, School of Literature, Media, and Communication, Georgia Tech. Dr. Klein is interested in the impact of this project on ideas about nineteenth-century women’s political writing. She will oversee the selection of the texts, and she will assess the accuracy of the social networks on the basis of her knowledge of the figures involved.

**Janet Murray (Advisor).** Associate Dean for Research and Faculty Affairs, Ivan Allen College of Liberal Arts, Georgia Tech. Dr. Murray has a long history of grant-funded work that bridges the humanities and computer science. She will advise the team on issues of interdisciplinary research.

**John Stasko (Advisor).** Professor, School of Interactive Computing, Georgia Tech. Dr. Stasko is a leading scholar in the field of information visualization. He will advise the team on issues relating to his specialty, as well as more general human-computer interaction concerns.

We plan to fund **one graduate student** who will implement both the underlying computational system and the user interface. Georgia Tech has strong doctoral programs in both Digital Media and Computer Science.

### Final product and dissemination

We plan to publish the results of both aspects of our initial analysis: the methods in a text-mining or visualization venue such as *Computational Linguistics* or *Information Visualization*, and the humanistic findings in a leading journal of American literary and cultural criticism, such as *American Literary History*; we may also publish the results of our collaborative research in a digital humanities journal. In addition, we will document the progress of the project on the research blog of the Georgia Tech Digital Humanities Lab, run by Dr. Klein.

The final product will be made publicly-available as an interactive web application, hosted on web servers maintained by Georgia Tech’s College of Computing. We will publicize the tool through various outlets, including the *Digital Humanities* conference. Our white paper will describe the technical and humanistic findings, as well as the issues we encountered in our attempt to investigate new computational methods for literary analysis through direct cross-disciplinary collaboration.

## 5. Project budget

### Category 1: Salaries and Wages

Project Director (Lauren Klein)	(b) (6)
Project Co-Director (Jacob Eistenstein)	(b) (6)
PhD Student	(b) (6)

### Category 2: Fringe Benefits

Senior personnel @ 27.9%	(b) (6)
Graduate research assistant @ .018%	(b) (6)

### Category 4: Travel

To Directors' Meeting (PD and CPD)	\$842
To <i>Digital Humanities</i> Conference (PD)	\$1022

### Category 7: Other Costs

Graduate Student Remission	\$10,530
----------------------------	----------

### TOTAL DIRECT COSTS

**\$47,174**

### Category 9: Total Indirect Costs

@ negotiated reduced overhead rate of 35%	\$12,825
-------------------------------------------	----------

### TOTAL INDIRECT COSTS

**\$12,825**

### TOTAL PROJECT COSTS

**\$59,999**

### Category 11: Project Funding

Requested from NEH	
Outright	\$59,999
Matching Funds	\$0
Subtotal	\$59,999
Cost sharing	\$0

### TOTAL PROJECT FUNDING

**\$59,999**

### ***Budget Narrative***

The grant funds will be primarily used to support one PhD student at 45% (full time), at the Georgia Tech College of Computing rate of (b) (6)/month, for the two academic semesters during which she will implement the topic model and interface components.

(b) (6) has been allocated as summer salary for the Project Director and Project Co-Director (1/4 of total summer salary).

In accordance with University policy, an additional (b) (6) has been added for Fringe Benefits at a rate of 27.9% for senior personnel and .018% for graduate research assistance. A total of \$12,825 has been allocated for Indirect Costs, at a reduced overhead rate of 35%, as negotiated with Georgia Tech Office of Sponsored Programs on 9/20/12.

A total of \$1864 has been allocated for travel. This includes \$842 to allow the Project Director and Co-Director to travel to the required Directors' Meeting, in Washington DC, calculated as follows:

\$400 plane fare (2 x \$200)  
\$300 for lodging (2 x \$150)  
\$142 daily stipend (2 x \$71, federal daily rate)

The remaining \$1022 has been allocated for the Project Director to travel to the *Digital Humanities* conference at the University of Nebraska, Lincoln, in July 2013, calculated as follows:

\$530 plane fare  
\$308 lodging (4 x \$77, federal daily rate)  
\$184 daily stipend (4 x \$46, federal daily rate)

Per University policy, tuition remission at the rate of \$1,170 per month in year one totaling \$10,530 has been allocated for all of the graduate assistants.

## 6. Biographies

### **Jacob Eisenstein**

Dr. Jacob Eisenstein is an Assistant Professor in the School of Interactive Computing at Georgia Tech. He works on statistical natural language processing, focusing on social media analysis, discourse, and non-verbal communication. The author of several dozen refereed papers, his research has been funded by the National Science Foundation and Google. Dr. Eisenstein was a Postdoctoral researcher at Carnegie Mellon and the University of Illinois. He completed his Ph.D. at MIT in 2008, winning the George M. Sprowls dissertation award.

### **Lauren Klein**

Dr. Lauren Klein is an Assistant Professor in the School of Literature, Media, and Communication at Georgia Tech. Her research interests include early American literature and culture, food studies, and the digital humanities. Founder of the Georgia Tech Digital Humanities Lab, her writing has appeared in *Early American Literature*, *American Quarterly*, and *In Media Res*. She received her A.B. from Harvard University and her Ph.D. from the City University of New York. Between 2007 and 2008, she worked as an educational technology consultant for One Laptop per Child, a non-profit aimed at bringing low-costs laptops to children in the developing world.

### **Janet Murray**

Dr. Janet Murray is Associate Dean for Research and Faculty Affairs and Ivan Allen College Dean's Recognition Professor in the School of Literature, Media, and Communication. She received her Ph.D. in English from Harvard University. Her primary research interests are interactive design, interactive narrative, and the history and development of representational media. She is the author of *Hamlet on the Holodeck: The Future of Narrative in Cyberspace* (MIT Press, 1998), and *Inventing the Medium: Principles of Interaction Design as a Cultural Practices* (MIT Press, 2011). Her eTV group creates prototypes of advanced broadband applications (<http://etv.gatech.edu>). She is also working on projects focused on engineering education (<http://intel.gatech.edu>) and on the elaboration of narrative schema for multisequential storytelling.

### **John Stasko**

Dr. John Stasko is Professor and Associate Chair of the School of Interactive Computing at Georgia Tech. He received his B.S. from Bucknell University, and his Sc.M. and Ph.D. from Brown University. His primary research area is human-computer interaction. He is the Director of the Information Interfaces Research Group. One central focus of a number of their projects is the creation of Information Visualization tools to help people understand large data sets. Another project focus is on evaluating anthropomorphic software agents/characters that are used as aids or filters in user interfaces.

## 7. Data Management Plan

### Data to be Generated

Type of Data	When shared?	Under what conditions?
Open Source computer code associated with tool, including algorithms and interface.	At conclusion of the start-up project, when initial testing has been completed.	Code will be freely available.
Dataset from ProQuest “Black Abolitionist Papers” database, to be used as test data.	ProQuest materials; cannot be shared. See next item.	N/A.
Dataset of freely-available abolitionist newspapers, to be released with computer code.	At conclusion of the start-up project, when initial testing has been completed.	Data will be freely available; see “Appendix D” for title list and sources.
Multimedia progress reports.	At the time of their writing, throughout the duration of the project.	The progress reports will be freely available on the Digital Humanities Lab website.
White paper.	After the project has been completed.	The white paper will be freely available on the Digital Humanities Lab website.
Multimedia report posted on the Digital Humanities Lab blog.	After the project has been completed.	The multimedia report will be freely available on the Digital Humanities Lab website.
Final report to NEH.	At the conclusion of the project.	Dissemination of the final report will be the responsibility of the NEH.

### Period of Data Retention

Data will be retained for 5 years beyond the completion of the start-up phase of TOME. Formal reports will be publically available within 1 year of project completion, on the Digital Humanities Lab website and on SMARTech (<http://smartech.gatech.edu/>), the Georgia Tech DSpace repository. Copies of the data will also be stored long-term on SMARTech.

### Data Formats and Dissemination

Computer code (Java and JavaScript) and public dataset (plain text) will be available as open source on SMARTech and GitHub, a publically-accessible code repository. All metadata associated with test dataset files will also be made available through these venues. Reports will be made available in PDF format and disseminated via the Digital Humanities Lab website (<http://dhlab.lmc.gatech.edu>).

### Data Management and Maintenance

All computer code and public test data will be stored in Github, where Eisenstein’s previous projects have been stored. All other data, including test data that cannot be shared, as well as any reports and publications, will be stored in SMARTech, Georgia Tech’s online institutional repository for faculty and researchers.

SMARTech is part of the MetaArchive Cooperative distributed preservation network, a large-scale effort for the preservation of electronic scholarly materials through the Library of Congress’s National Digital Information Infrastructure and Preservation Program (NDIIPP). For the details of its preservation methods, see: <http://smartech.gatech.edu/policy>.

## 8. Letters of Support

Please see the following pages for letters of support from:

**Dr. Amy Earhart**, Assistant Professor  
Department of English  
Texas A&M University  
Phone: (979) 862-3038  
Email: aearhart@tamu.edu

**Dr. David Mimno**, CRA Computing Innovation Fellow  
Department of Computer Science  
Princeton University  
Phone: (609) 258-9907  
Email: mimno@cs.princeton.

September 12, 2012

Professor Lauren Klein  
School of Literature, Media, and Communication  
Georgia Tech  
781 Marietta Street  
Atlanta, GA 30332-0525

Dear Professor Klein:

I am very pleased to write a letter of support for TOME: Interactive Topic Model and Metadata Visualization. The recent scholarly interest in data mining and visualization represents an exciting moment for digital scholarship, and TOME is a welcome addition to the field. I have known Lauren Klein and her work for four years. I remain consistently impressed with her ability to merge algorithmic and literary approaches to texts, as well as her consistent care for issues of race and gender. Her work on Thomas Jefferson and James Hemings is invaluable, and I use the project in my undergraduate and graduate classroom to discuss best practices in data mining and visualization. Lauren's work represents the best of the new approaches, as she consistently attends to a complex set of traditional literary questions while using innovative technologies to broaden our understanding of a text.

TOME is an exciting proposal. Expanding current work in topic modeling, TOME promises to allow scholars to begin to reveal the way ideas circulated among social and literary movements. The four research questions outlined by the proposal—the what, who, when and how—have not been adequately answered using traditional humanistic approaches to large sets of data. While tools such as WordSeer and MONK are incredibly useful for mining bodies of texts, they are not as accurate for use in earlier textual materials, which may display different syntactic patterns than the contemporary data sets from which the tools are drawing comparisons. Varying dialect further complicates algorithmic applications and, as would be the case with the abolitionist texts proposed in the sample set of the TOME proposal, the accuracy further drops. TOME offers a way to conduct the type of broad comparisons imagined by those interested in data mining on texts that are often ignored or underrepresented, texts that include dialect and non-standard English.

TOME's proposed visualization will allow the scholar more control than current visualizations. Indeed TOME is designed to work as a tool by which the scholar might more carefully control and manipulate the results of the algorithm. Current visualization tools display results in a fairly flat form and results are often difficult for users to manipulate. TOME promises 3D visualizations that the user might fine tune and revisualize. Such a tool would be immensely useful and a step forward in visualizations of textual data.

227 John R. Blocker Building  
4227 TAMU  
College Station, TX 77843-4227

Tel. 979.845.3452 Fax. 979.862.2292  
[www-english@tamu.edu](http://www-english@tamu.edu)

Selfishly, I'm thrilled to learn that the team has chosen the Black Abolitionist Papers as their test set of data. I have spent a good deal of time working with these papers and the question of how ideas circulate, as well as the progression of ideas related to activism, are still open questions. My current work has focused on the examination of Malcolm X's ideology in transition, and TOME would allow me to connect Malcolm X's work to the broader set of black nationalist thought over a broad period. Even more importantly, the tool is designed to prove applicable to issues of race and gender, at times almost impossible to accurately represent with current tools as previously discussed.

Best wishes for the success of TOME. I strongly recommend that the NEH Digital Start-Up grant be funded.

Sincerely,

A handwritten signature in black ink, appearing to read 'Amy Earhart', followed by a large, stylized circular flourish or scribble.

Amy E. Earhart

David Mimno  
Department of Computer Science  
Princeton University  
35 Olden St.  
Princeton, NJ 08540

Sept 22, 2012

Brett Bobley  
Office of Digital Humanities  
National Endowment for the Humanities

Dear Mr. Bobley,

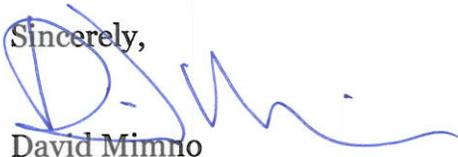
I am writing to support the Georgia Tech proposal "TOME" to the Digital Humanities Startup program. This project will provide needed tools and methodologies for the analysis of large collections of historical documents. The focus on networks of influence and author characteristics such as gender will be significant contributions.

As the author of the MALLETT topic modeling package, I can say that there is no one more qualified to guide the technical side of this project than Dr. Eisenstein. I believe that a collaboration between Dr. Klein and Dr. Eisenstein will benefit both humanities scholarship and computer science research.

A particular strength of the proposal is the focus on a data analysis *lifecycle*. It is tempting to think of topic modeling and other types of statistical analysis as simply a process that can be run. But topic modeling is not yet easy to use "out of the box". Practitioners commonly find that many iterations of vocabulary selection and corpus refinement are necessary to extract useful topics. I am especially happy therefore to see the proposal's emphasis on iterative approaches that allow scholars to generate topics and queries that are useful for their specific needs.

If digital technologies are to become a vital part of humanities research, scholars will need methodologies that are standard enough to be widely accepted yet flexible enough to accommodate a wide variety of applications. This project has the potential to provide both digital tools, but also --- and more importantly --- processes for conducting research using large text collections.

Sincerely,



David Mimno

## 9. Appendices

### Appendix A: Topic Modeling Background

The primary computational technique that underlies our project is topic modeling. Topic modeling is a text mining technique that applies probabilistic inference to identify latent themes, or “topics,” in a set of documents. The documents themselves serve as the topic model’s input. The model’s output consists of two parts: a set of topics—that is, clusters of words that appear in similar documents—and the topical composition of each document. For example, our research on the *Papers of Thomas Jefferson* revealed the following topics (selected from 30 topics in total):

- defender, writ, sheriff, plaintiff, penalty, testator, damages
- maria, she, daughter, papa, her, sister, polly, love, martha
- congress, senate, amendments, vote, confederation, resolutions, majority

Each topic clearly indicates a different theme in the Jefferson archive: legal issues, family matters, and political concerns. Topics such as these, here summarized by short phrases, constitute the first portion of any topic model’s output.

The second portion of the model’s output consists of a characterization of each document in terms of the topics it contains. On the basis of the topics described above, one document might be characterized as 10% legal issues, 45% family matters, and 45% political concerns; another might be characterized as 80% legal issues, 5% family matters, and 15% political concerns. It is important to underscore that these themes are not defined in advance, but rather are extracted from the set of documents—or archive—provided to the topic model.

Topic models are based on statistical inference over a generative model of the archive – a probabilistic explanation for how the archive was produced. The generative model works as follows: each document is assumed to have an hidden vector of topic strengths, which summarizes the main themes in the document. Next, to generate each word in the document, the author works one word at a time: first choosing a topic (with probability equal to the topic strength in the document), and then selecting a word from the topic (which assigns some probability to each word in the vocabulary).

This describes the modeling assumptions, but in practice we observe only the words, and must work backwards through statistical inference to recover the topic-word distributions and the topic strengths for each document. The output from this process is a set of topics (probability distributions over words), and the hidden topic strengths for each document. Each topic is typically summarized by the words which are much more probable in the topic than they are in the corpus overall. The tutorial by Blei (2012) provides more technical detail.

Generic topic models are computed from words alone, without consideration for metadata such as author or publication date. However, by tracing the relationship between topics and metadata, a topic model visualization can provide an entry point for more precise scholarly work. It is also possible to incorporate metadata into the topic modeling process, which helps to ensure that the extracted themes cohere with the metadata of interest (Rosen-Zvi et al. 2004; Mimno and McCallum 2006).

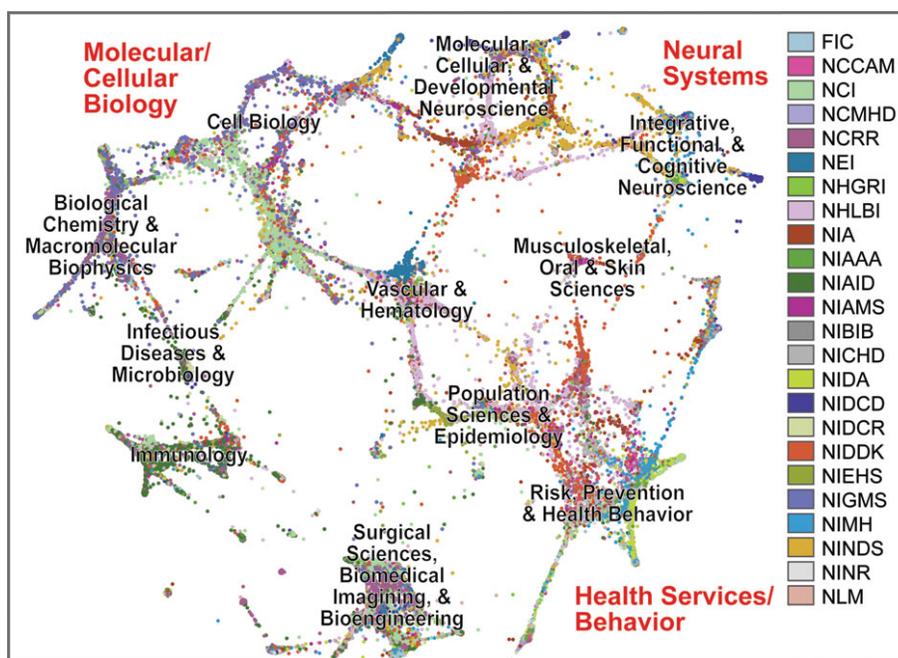
### Appendix B: Modeling Influence

Modeling influence and causality is an unsettled problem in statistics, though there are approaches which seem potentially relevant (for example, see Nallapati and Manning 2011). Our basic assumption is that topic-specific influence is characterized by at least two properties. First,

the author must write about the topic to an unusually large extent over some timespan. Second, other authors must increase the extent to which they write about the same topic in later times. Granger causality proposes that causal influence can be detected between  $i$  and  $j$  if information about  $i$  at time  $t$  helps to improve the predictions of  $j$  at some later time (Granger 1969). We can estimate Granger causality by fitting autoregressive models to the estimated topic proportions, and then determine which sets of authors exert influence on each other. The survey by Guyon et al. (2008) provides more detail, and summarizes recent approaches.

### Appendix C: Text Visualization

The problem of text visualization has been considered since at least the mid-1990s, with early work focusing on projections and similarity-based layouts. For example, the Galaxies visualization of Wise et al. (1995) represents documents as dots in a spatial map, with inter-document distance set by the similarity in word frequencies between each pair of documents. Documents that use many similar words will be near each other, while documents with very different content will be distant. This idea was later extended to visualize the output of topic models by Talley et al. (2011). In the map below, documents (in this case, the archive consists of NIH grant proposals) are located near each other based on the similarity of their topic representations. This produces a “map” of the archive. The authors have then provided their own names for clusters in the map.



Such maps provide an immediate understanding of the thematic structure of the archive, but by forcing all thematic differences into a single two dimensional presentation, information is inevitably lost. TOME takes an alternative approach, dynamically producing new visualizations for subsets of documents, and allowing the user to focus on the local thematic landscape of a the most relevant topics and documents.

**Appendix D: Public Dataset**

Because our test dataset will derive from the pay-for-access ProQuest “Black Abolitionist Papers” database, to which Georgia Tech subscribes, but many individuals and institutions do not, we have also begun to construct a smaller public dataset consisting of significant titles in the Proquest database that have also been digitized and made freely available through other means. The table below indicates the titles and sources of several such newspapers. We expect this list to continue to expand as our project develops:

<b>Newspaper Title</b>	<b>Accessible Archive</b>	<b>LoC Historic American Newspapers</b>	<b>Many Roads to Freedom</b>	<b>Other</b>
Anti-Slavery Bugle (1845-61)		x		
Christian Recorder (1854-92)	x			
Colored American (1837-41)	x			
Frederick Douglass’s Paper (1851-63)	x		x	
Freedom’s Journal (1827-29)	x			Cornell U. Library
Liberator (1831-65)				Liberator Files
National Era (1847-60)	x			Google News
New-York Tribune (1841-1966)		x		
Northern Freeman (1848)			x	
North Star (1847-51)	x		x	
Provincial Freeman (1854-57)	x			
Rites of Man (1834)			x	
Voice of the Fugitive (1851-52)			x	
Weekly Advocate (1837)	x			

## Appendix E: Works Cited

- Blei, D. "Probabilistic topic models." Communications of the ACM. 55.4 (2012): 77–84.
- Blevins, C. "Topic Modeling Historical Sources: Analyzing the Diary of Martha Ballard" Proceedings of Digital Humanities. 2011.
- Davis, D.B. Antebellum American Culture: An Interpretive Anthology. University Park: Penn State UP, 1997 [1979].
- Dudden, F. Fighting Chance: The Struggle Over Woman Suffrage and Black Suffrage in America. New York: Oxford UP, 2011.
- Eisenstein, J., D. H. Chau, A. Kittur, E. P. Xing. "TopicViz: Semantic Navigation of Document Collections." Supplemental Proceedings of Conference on Human Factors in Computing Systems (CHI). 2012.
- Gelman, A. "Causality and Statistical Learning." American Journal of Sociology. 117.3 (2011): 955-966.
- Gildea, D. "Corpus Variation and Parser Performance." Proceedings of Empirical Methods in Natural Language Processing. 2001.
- Jockers, M. Macroanalysis: Digital Methods and Literary History. U Illinois P, 2013).
- Karcher, C. The First Woman in the Republic: A Cultural Biography of Lydia Maria Child. Durham: Duke UP, 1997 [1994].
- McCallum, A. K. MALLET: A Machine Learning for Language Toolkit. <http://mallet.cs.umass.edu>. 2002
- Mei, Q., X. Shen, and C. Zhai. "Automatic labeling of multinomial topic models." Proceedings of Conference on Knowledge Discovery and Data Mining (KDD). 2007. 490-499.
- Mimno, D. "Computational historiography: Data mining in a century of classics journals." Journal on Computing and Cultural Heritage. (2012)
- Mimno, D., M. Hoffman, and D. Blei. "Sparse stochastic inference for latent Dirichlet allocation." Proceedings of International Conference on Machine Learning. 2012.
- Mimno, D. and A. McCallum. "Topic models conditioned on arbitrary features with dirichlet-multinomial regression." Proceedings of Conference on Uncertainty in Artificial Intelligence. 2008.
- Nallapati, R. and C. Manning. "TopicFlow model: Unsupervised learning of topic specific influences of hyperlinked documents." Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS). 2011.
- Peterson, C. Black Gotham: A Family History of African Americans in Nineteenth-Century New York. New Haven: Yale UP, 2011.

Rhody, L. "Some Assembly Required: Understanding and Interpreting Topics in LDA Models of Figurative Language." Lisa @ Work. 2012.

Rosen-Zvi, M., T. Griffiths, M. Steyvers, and P. Smyth. "The author-topic model for authors and documents." Proceedings of the 20th conference on Uncertainty in artificial intelligence. 2004.

Sekine, S. "The Domain Dependence of Parsing." Proceedings of 5th Conference on Applied Natural Language Processing. 1997.

Stanstell, C. The Feminist Promise: 1792 to the Present. New York: Modern Library, 2010.

Talley, E. M., D. Newman, D. Mimno, B. W. Herr, H. M. Wallach, G. A. P. C. Burns, A. G. M. Leenders, and A. McCallum. "Database of NIH grants using machine-learned categories and graphical clustering." Nature Methods, 8(6):443–444, May 2011

Underwood, T. "The Differentiation of Literary and Nonliterary Diction, 1700-1900." The Stone and the Shell. 2011.

Yang, T., A. J. Torget, and R. Mihalcea, "Topic Modeling on Historical Newspapers." Proceedings of the Association for Computational Linguistics workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (ACL LATECH 2011), June 2011. 96-104

Yi, J. S., R. Melton, J. Stasko, and J. A. Jacko. "Dust & magnet: multivariate information visualization using a magnet metaphor." Information Visualization, 4:239–256, 2005.