OMB Number: 4040-0003
Expiration Date: 02/28/2022

APPLICATION FOR FEDERAL DOMESTIC ASSISTANCE - Short Organizational

*** 1. NAME OF FEDERAL AGENCY:**

National Endowment for the Humanities

2. CATALOG OF FEDERAL DOMESTIC ASSISTANCE NUMBER:

45.149

CFDA TITLE:

Promotion of the Humanities Division of Preservation and Access

*** 3. DATE RECEIVED:** Completed Upon Submission to Grants.gov **SYSTEM USE ONLY**

*** 4. FUNDING OPPORTUNITY NUMBER:**

20190515-PR

* TITLE:

Research and Development

5. APPLICANT INFORMATION

* a. Legal Name:

Board of Trustees of the University of Illinois

b. Address:

* Street1: 1901 S. First Street, Suite A	Street2:
* City: Champaign	County/Parish:
* State: IL: Illinois	Province:
* Country: USA: UNITED STATES	* Zip/Postal Code: 61820-7406

c. Web Address:

http://

* d. Type of Applicant: Select Applicant Type Code(s): H: Public/State Controlled Institution of Higher Educ	* e. Employer/Taxpayer Identification Number (EIN/TIN): 376000511
Type of Applicant:	* f. Organizational DUNS: 0415440810000
Type of Applicant:	* g. Congressional District of Applicant: IL-013
* Other (specify):	

6. PROJECT INFORMATION

* a. Project Title:

Broadening Access to Text Analysis by Describing Uncertainty

* b. Project Description:

The noise associated with digital transcription has become an important obstacle to humanistic research. While the errors in digital texts are easily observed, the downstream effects of error on scholarship are far from clear. Consequential problems for the humanities often spring less from the average level of error in a collection than from the uneven distribution of noise across different periods, genres, and social strata. Uncertainty about this problem undermines confidence in research and discourages some scholars from using digital libraries at all. To address these problems, we will 1) Create paired libraries of clean, manually transcribed volumes and optically-transcribed versions of the same volumes, with or without paratext. 2) Conduct parallel experiments in these corpora to empirically measure the distortions affecting scholarship. 3) Construct a map of error and share resources that help scholars estimate levels of uncertainty in their work.

c. Proposed Project: * Start Date: 03/01/2020 * End Date: 05/31/2021

APPLICATION FOR FEDERAL DOMESTIC ASSISTANCE - Short Organizational

7. PROJECT DIRECTOR

Prefix: Prof.	* First Name: William
Middle Name:	
* Last Name: Underwood	Suffix:
* Title: Professor of English and Information Sciences	* Email: tunder@illinois.edu
* Telephone Number: 217-244-4617	Fax Number:
* Street1: 501 E. Daniel Street	Street2:
* City: Champaign	County/Parish:
* State: IL: Illinois	Province:
* Country: USA: UNITED STATES	* Zip/Postal Code: 618200000

8. PRIMARY CONTACT/GRANTS ADMINISTRATOR

☐ Same as Project Director (skip to item 9):

Prefix:	* First Name: Robin
Middle Name:	
* Last Name: Beach	Suffix:
* Title: Director, Pre-Award	* Email: spa@illinois.edu
* Telephone Number: 217-333-2187	Fax Number:
* Street1: 1901 S. First Street, Suite A	Street2:
* City: Champaign	County/Parish:
* State: IL: Illinois	Province:
* Country: USA: UNITED STATES	* Zip/Postal Code: 61820-7406

APPLICATION FOR FEDERAL DOMESTIC ASSISTANCE - Short Organizational

9. * By signing this application, I certify (1) to the statements contained in the list of certifications** and (2) that the statements herein are true, complete and accurate to the best of my knowledge. I also provide the required assurances** and agree to comply with any resulting terms if I accept an award. I am aware that any false, fictitious, or fraudulent statements or claims may subject me to criminal, civil, or administrative penalties (U.S. Code, Title 218, Section 1001)

** I Agree ☒

** The list of certifications and assurances, or an internet site where you may obtain this list, is contained in the announcement or agency specific instructions.

AUTHORIZED REPRESENTATIVE

Prefix:	* First Name:	Middle Name:
	Avijit	

* Last Name:	Suffix:
Ghosh	

* Title:	* Email:
Comptroller	spa@illinois.edu

* Telephone Number:	Fax Number:
217-333-2187	

* Signature of Authorized Representative:	* Date Signed:
Completed by Grants.gov upon submission.	Completed by Grants.gov upon submission.

OMB Number: 3136-0134
Expiration Date: 06/30/2021

Supplementary Cover Sheet for NEH Grant Programs

1. Project Director Major Field of Study: Interdisciplinary: General

2. Institution Information Type: 1330: University

3. Project Funding

Outright Funds	73,122.00
Federal Match	0.00
Total from NEH	73,122.00
Cost Sharing	0.00
Total Project Costs	73,122.00

4. Application Information

Will this proposal be submitted to another NEH division, government agency, or private entity for funding?

☐ Yes
☒ No

If yes, please explain where and when:

Type of Application

☒ New
☐ Supplement

If supplement, list current grant number(s).

Primary project discipline	Interdisciplinary: General
Secondary project discipline (optional)	Literature: British Literature
Tertiary project discipline (optional)	Literature: American Literature

Project/Performance Site Location(s)

Project/Performance Site Primary Location

☐ I am submitting an application as an individual, and not on behalf of a company, state, local or tribal government, academia, or other type of organization.

Organization Name: Board of Trustees of the University of Illinois

DUNS Number: 0415440810000

* Street1: 506 S. Wright Street

Street2:

* City: Urbana

County:

* State: IL: Illinois

Province:

* Country: USA: UNITED STATES

* ZIP / Postal Code: 61801-3620

* Project/ Performance Site Congressional District: IL-013

Project/Performance Site Location 1

☐ I am submitting an application as an individual, and not on behalf of a company, state, local or tribal government, academia, or other type of organization.

Organization Name:

DUNS Number:

* Street1:

Street2:

* City:

County:

* State:

Province:

* Country: USA: UNITED STATES

* ZIP / Postal Code:

* Project/ Performance Site Congressional District:

Additional Location(s)

Add Attachment | Delete Attachment | View Attachment

ATTACHMENTS FORM

Instructions: On this form, you will attach the various files that make up your grant application. Please consult with the appropriate Agency Guidelines for more information about each needed file. Please remember that any files you attach must be in the document format and named as specified in the Guidelines.

Important: Please attach your files in the proper sequence. See the appropriate Agency Guidelines for details.

1) Please attach Attachment 1	abstract.pdf	Add Attachment	Delete Attachment	View Attachment
2) Please attach Attachment 2	contents.pdf	Add Attachment	Delete Attachment	View Attachment
3) Please attach Attachment 3	narrative.pdf	Add Attachment	Delete Attachment	View Attachment
4) Please attach Attachment 4	budget.pdf	Add Attachment	Delete Attachment	View Attachment
5) Please attach Attachment 5	appendices.pdf	Add Attachment	Delete Attachment	View Attachment
6) Please attach Attachment 6	particiantslist.pdf	Add Attachment	Delete Attachment	View Attachment
7) Please attach Attachment 7	letters.pdf	Add Attachment	Delete Attachment	View Attachment
8) Please attach Attachment 8	dissemination.pdf	Add Attachment	Delete Attachment	View Attachment
9) Please attach Attachment 9	agreement.pdf	Add Attachment	Delete Attachment	View Attachment
10) Please attach Attachment 10		Add Attachment	Delete Attachment	View Attachment
11) Please attach Attachment 11		Add Attachment	Delete Attachment	View Attachment
12) Please attach Attachment 12		Add Attachment	Delete Attachment	View Attachment
13) Please attach Attachment 13		Add Attachment	Delete Attachment	View Attachment
14) Please attach Attachment 14		Add Attachment	Delete Attachment	View Attachment
15) Please attach Attachment 15		Add Attachment	Delete Attachment	View Attachment

Broadening Access to Text Analysis by Describing Uncertainty

A Tier I NEH Research and Development Proposal

1. Abstract.

At the end of the twentieth century, humanists envisioned text mining primarily as a problem of information retrieval. But scholars now commonly use digitized texts to measure historical or literary trends across a long timeline (Armitage and Guldi; Underwood).

This expanding research program puts new pressures on digital libraries. Most of the text in our largest digital libraries is optically transcribed, and optical transcription creates error. While the effect of error on particular algorithms has been studied, its downstream consequences for the historical conclusions humanists actually want to draw are not yet well understood. In particular, many existing studies assume that error is randomly distributed. This was a fair assumption for some information retrieval applications, but it becomes importantly false for the comparative historical questions contemporary scholars are exploring. Unevenly distributed errors can distort humanistic conclusions in significant and unexpected ways.

As a result of this uncertainty, many scholars avoid using large digital libraries altogether, relying instead on smaller (often proprietary) collections with cleaner text. A Mellon-funded report exploring this problem has recommended that, to encourage and guide appropriate use of digital libraries, "researchers should survey and perform quantitative evaluations of the effect of OCR error on commonly used text-analysis methods" (Smith and Cordell 14).

This proposal responds directly to that suggestion by undertaking an empirical survey of the consequences of error (and paratextual noise) for the historical conclusions humanists actually build on text analysis. We focus on English-language contexts, but also consider Chinese-language text mining as a comparative touchstone. The project will enhance access to digital libraries in three ways:

- We will create a map of error and paratextual noise in English-language books from 1700 to 2010.
- We will create a library of at least a thousand English-language texts, and a hundred Chinese-language texts, each of which exists in manually and optically transcribed versions, with or without paratext. By running parallel experiments in different versions of this dataset, researchers will be able to assess the distorting effects of noise on their specific research questions.
- We will distribute Jupyter notebooks that help researchers estimate the downstream uncertainty associated with specific levels of error in different research methods.

Finally, we will summarize our conclusions in peer-reviewed articles, and propose a plan for developing a library of "work-level" body texts separated from paratext.

2. Table of contents.

3. Project narrative.

a. Significance.

Questions about the reliability of digital libraries have hovered over the humanities for several decades, growing in urgency as scholars rely more heavily on digitized text. In 2017, those concerns moved to the center of the national agenda when the Andrew W. Mellon Foundation funded a project at Northeastern University "to study the current state of optical character recognition (OCR) for historical and multilingual documents and to outline future directions for research in this area" (Quinn). A report appeared in 2018, and the NEH has encouraged researchers to frame proposals that respond to its recommendations. This proposal aims to foster wider use of digital collections by responding directly to the report's first recommendation: that "researchers should survey and perform quantitative evaluations of the effect of OCR error on commonly used text-analysis methods" (Smith and Cordell 14).

As Smith and Cordell point out, lack of clarity about the consequences of error leads many scholars to avoid using open digital libraries entirely. They may rely instead on smaller, proprietary collections (10). This is a barrier to access, and not an easy one to remove. While it can be relatively easy to quantify error itself, it is unsafe to assume that the conclusions reached in text analysis are distorted in proportion to the mean level of error in a collection. On the contrary, the effects of error can vary enormously from one method or question to another. In some forms of natural language processing, a single mistranscribed letter may vitiate a whole sentence. Other methods are barely disturbed at all by lexical errors, as long as those errors are equally distributed across texts. But are errors in fact distributed equally across authors, genres, and periods? This crucial dimension of the problem has so far received little attention. Researchers may also be falsely reassured by a definition of reliability that focuses too narrowly on OCR accuracy. In the real world, digitization can create other pitfalls: for instance, a computer doesn't necessarily know how to separate the text of a novel from a biographical introduction, or from the running headers at the top of each page.

In short, questions about the reliability of digital collections cannot be answered simply by identifying an acceptable mean level of transcription error. This is rather an empirical question about the interaction of different research questions with the actual, uneven distribution of noise in existing corpora. If we understood that interaction, researchers could choose different corpora or different methods. They might undertake more ambitious projects, and they could certainly do a better job of communicating the uncertainty associated with their findings. Finally, funding agencies might be guided toward more consequential correction strategies.

We propose to advance these goals by mapping the consequences of error and paratextual noise for real-world humanistic questions across different languages (English and Chinese), periods (1700 to the present), and analytical methods (including emerging neural approaches). We will also evaluate several different correction strategies, including post-processing correction based on language modeling and the collation of variant copies. While a Tier I grant has a relatively limited budget, our reach will be extended by collaboration with the Ichneumon project, directed by David A. Smith, and with HathiTrust Research Center. The

deliverables of the project will include a report that explains our conclusions, a dataset that maps error in HathiTrust, and a set of Jupyter notebooks that help researchers estimate the level of uncertainty associated with their conclusions given the actual distribution of noise in their documents. We will also share a library of at least 1,000 English-language volumes 1700-1924, each of which is available in five different versions (ranging from clean, manually transcribed ground truth to optically transcribed versions with or without paratext). By running parallel experiments in different versions of this corpus, researchers will be able to estimate the distorting effect of noise on their own research question. Finally, we will develop prototypes and plans for a more ambitious project to separate paratext from body text in digital libraries.

b. Background of applicant.

Ted Underwood, the PI, is appointed both in Information Sciences and in English at the University of Illinois, Urbana-Champaign, and will draw on resources associated with both disciplines. Illinois' School of Information Science is a co-host (with Indiana University) of HathiTrust Research Center. Collaboration with HTRC will guarantee access to optically transcribed books even in the twentieth and twenty-first centuries; computational resources for the project are available through Illinois' Campus Cluster.

But the key differentiating feature of this project is that it proposes, not to measure noise, but to assess its actual consequences for the comparative questions humanists are using text analysis to pose. It is thus essential to the project's success that it be led by someone who has experience basing substantive historical arguments on text analysis. The PI has this experience, with a track record of publications that show how text analysis can produce "specific discoveries that are redrawing our map of the last three hundred years of English-language literature" (Underwood).

c. History, scope, and duration.

C1. History of the project. The PI is cited in Smith and Cordell as one of "a handful of researchers working with errorful OCR data" (10). This experience has made him aware of important pitfalls that are left unmapped in existing scholarship.

Basic quantitative work has already been done on the relative vulnerabilities of different methods. Processes like part-of-speech tagging and entity extraction often degrade "linearly as a function of word error rates" (Lopresti 9). By contrast, bag-of-words approaches to text analysis can be surprisingly robust to error (Smith and Cordell 15; Franzini et al.; Eder). In a collection of eighteenth century books, for instance, topic modeling and authorial attribution were largely unaffected by OCR error until the fraction of accurately transcribed words sank below 80% (Hill and Hengchen). The reason, presumably, is that these algorithms are less concerned with the absolute frequency of any word than with the differences that separate groups of texts, and those *differences* may be largely unaffected by deformations that change texts in parallel ways.

But studies of this kind may give researchers misplaced confidence, because they typically consider questions (like authorial attribution) where noise is distributed more or less randomly

across the boundary of interest—degrading rather than distorting the signal. As text analysis is applied to larger historical questions, it becomes unsafe to assume that noise is random. Older books and cheaply printed books are more vulnerable to mistranscription.

Moreover, mistranscription may not be the most insidious problem created when pages designed for human eyes are flattened into a stream of tokens. A researcher studying the difference between fiction and biography would run into significant problems if they used whole digitized volumes, since works of fiction often begin with biographical prefaces. To estimate the significance of this problem, we constructed a sample of fifty books (both works of fiction and biographies) that were available in both clean (manually transcribed) and imperfect (optically transcribed) versions. We edited the clean text to include only the words a reader might consider the "body" of the text—excluding paratexts such as introductions, indexes, tables of contents, advertisements, due date slips, and running headers. Then we compared the clean texts to three versions of the imperfect text: a) the original version, warts and all, b) a version with OCR errors partially corrected using a rule-based system, but with paratext left in, and c) a version with OCR partially corrected and paratext trimmed out.

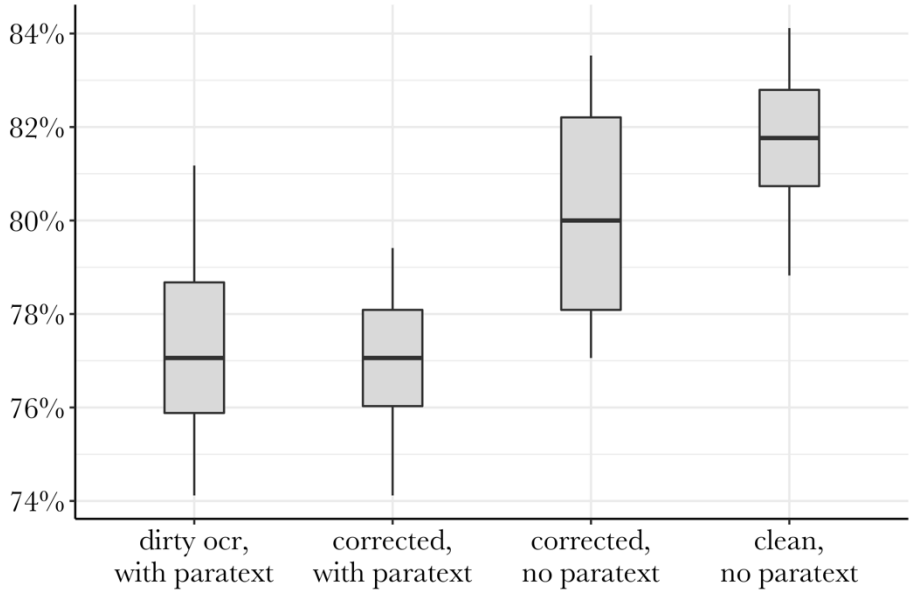


Figure 1. Accuracy of predictive models distinguishing fiction from biography, 1700-1900. All models are run on balanced samples of 85 texts from each class. Each "text" represents one fourth of a volume; we keep texts by the same author grouped in cross-validation.

When we constructed predictive models to distinguish biography from fiction, we found that the inclusion of paratext weakened model accuracy more than OCR error as such (Figure 1). While it is easy to see how paratext would specifically contaminate the boundary between fiction and biography, the same problem can lurk below the surface of other categories in a less obvious form. For instance, a researcher might imagine they were modeling subtle signals of literary

prestige in the texts of poems—while their model was actually relying on overt social signals given in introductions or in the prices of advertisements at the back of the book.

Noise can distort results in opposite directions depending on its distribution across the boundary of interest. In Figure 1, OCR error and paratextual noise makes a boundary between genres look blurrier than it truly is. But if we pose a different question about the same set of texts, noise may actually sharpen the boundary. In Figure 2, for instance, we have divided biographies and fiction chronologically, training models to distinguish volumes printed in the eighteenth and nineteenth centuries. Since OCR errors are common in eighteenth-century books, and significantly change word frequencies there, this chronological boundary is (perversely) easiest to model when OCR errors are left uncorrected. It becomes hardest to model with clean, manually transcribed documents. The high accuracy produced by dirty OCR might easily be misinterpreted as a sign of rapid linguistic change, when it actually signals only a change in print quality.

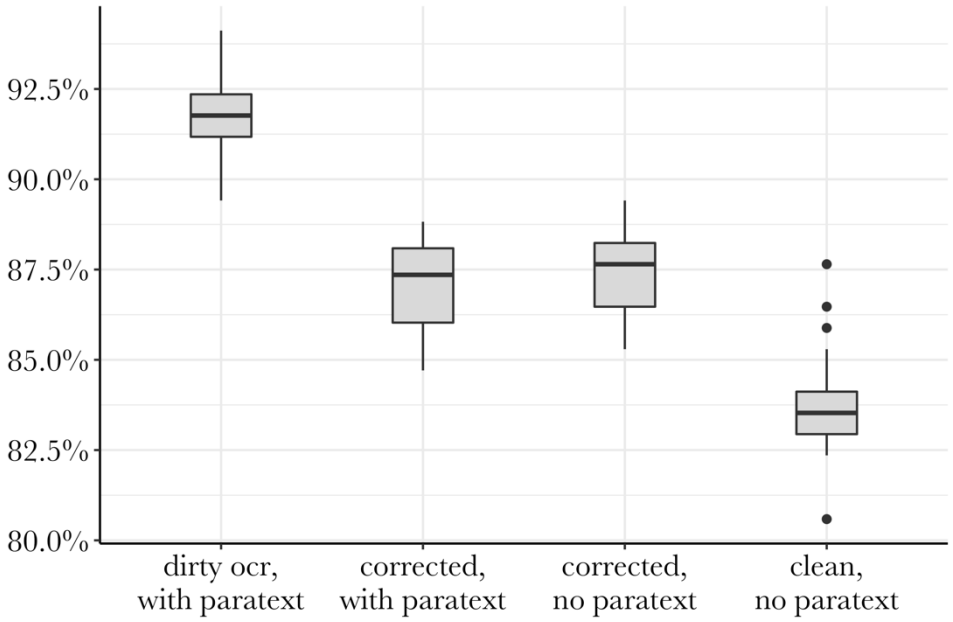


Figure 2. Accuracy of predictive models distinguishing books printed in the eighteenth and nineteenth centuries. All models are run on balanced samples of 85 texts from each class; books are divided and cross-validation is carried out as in Figure 1.

In short, existing studies don't give researchers enough guidance, because they focus too narrowly on transcription error (ignoring paratextual noise), and tend to consider questions where error is relatively random. Before digital libraries can provide a reliable foundation for text analysis, researchers need rules of thumb based on a wider sample of real-world research questions. Since no study can consider every possible scenario, a project of this kind should also create a larger library of paired texts resembling the small 50-volume sample we used to create Figures 1 and 2 above. By running experiments on corpora transcribed or corrected in different ways, with or without paratext, researchers could test the vulnerability of their own research question to various kinds of distortion.

In the sample above, front and back matter were identified manually, while running headers were trimmed with a fuzzy-matching algorithm. But more sophisticated strategies for identifying paratext are available, and will become necessary as the project scales up. In 2014, the PI of the current proposal produced a page-level map of paratext in HathiTrust up to 1923 (Underwood et al.) More recently, McConnaughey et al. have studied "The Labeled Segmentation of Printed Books," producing a large training dataset and recommendations for best practices. Finally, an NEH-funded project led by David A. Smith (Ichneumon) has the potential to produce an even larger pipeline of training data, by using sequence alignment to collate different editions of books, thereby highlighting the things (introductions, headers, etc) that change between editions (Smith). This practical definition of paratext as "the things that change" might prove more reliable than any abstract ontology, since (for instance) a glossary at the back is sometimes editorial paratext, and sometimes an inseparable part of a science fiction novel.

Ichneumon is also providing a new, state-of-the-art form of OCR correction, by collating multiple OCR transcripts of a text, and using the aligned texts to train correction models. We will use Ichneumon's texts wherever they are available, filling gaps with simpler correction methods where they aren't, and measuring OCR accuracy against manually transcribed ground truth to render different qualities of correction comparable.

C2. Scope. We propose to construct an English-language corpus of at least a thousand printed books, each of which is represented in at least five versions—one manually transcribed, paired with four optically transcribed versions filling out a 2x2 grid (with/without paratext, with/without partial OCR correction). These texts will be aligned at the page level.

We know a corpus of at least this size is achievable, because an experiment automatically matching HathiTrust English fiction against Gutenberg and TCP-ECCO metadata identified over a thousand titles present in both forms, 1700-1924. We expect to find an equal amount of nonfiction. Matched titles may not always represent the same edition, so we will manually check the collection to ensure that texts are comparable.

Volumes will be distributed across a timeline from 1700 to 1924, and across three broad genres (narrative fiction, poetry, and biography). We are focusing on specific genres, rather than randomly sampling printed books, because the most troubling questions about the reliability of text-mining involve comparative hypotheses. So we will need groups of books to compare.

Paratextual noise is also especially damaging for literary genres, since paratext tends to be very unlike body text there. So we have considered two literary genres, along with one nonfictional genre (biography). In this central corpus, we will test a wide range of methods and questions:

- Entity extraction, especially character extraction using BookNLP (Bamman et al.)
- Topic modeling. We will test the robustness of OCR models using KL divergence from models trained on a clean corpus.
- Predictive modeling of social boundaries, including
  - -- genre boundaries (fiction/biography)
  - -- authorial identities,
  - -- and, within fiction, strata of literary prestige as identified by book reviews.

  We will model these boundaries using
  - -- regularized logistic regression on bags of words,
  - -- but also a pre-trained transformer model, such as GPT, that requires sequential text. (The robustness of this method on noisy text from an older linguistic context is an important open question we aim to answer.)

We will also test some of these methods on a wider social context, so we can express our conclusions with an appropriate degree of caution. Using a HathiTrust Research Center data capsule, we will test entity extraction and neural classification on English-language volumes beyond the copyright cut-off at 1924. We will also construct a small Chinese-language corpus of at least a hundred volumes, each represented in optically- and manually-transcribed versions, in order to test the robustness of predictive modeling in a different linguistic context (where, for instance, word and phrase segmentation poses different problems than English-language tokenization).

Our goal in doing this is not to thoroughly address the challenges of Chinese-language text mining, but to qualify the conclusions we draw from an English context. Early experiments have led us to suspect that the identification of paratext is a larger problem for most real-world questions than the better-studied problem of OCR accuracy. If that hypothesis is borne out, we might write a report recommending further work to solve the problem. In that case, we want to be able to say whether our assessment of the relative importance of this problem applies only to Latin alphabets or to a larger slice of the library.

At this stage, we do not propose to extend our investigation to cover periodicals or newspapers. The document layout and segmentation problems in those media would add complications beyond the scope of a Tier I project.

C3. Duration. The project will extend for fourteen months, from March 1, 2020 to May 31, 2021.

d. Work products.

Aligned data sets. As described above, we will align manually and optically transcribed versions of at least 1,100 volumes, with and without paratext. We will share these datasets publicly so researchers can run experiments to assess the distortion created by transcription error and paratextual noise.

A map of estimated error. After training a model on these aligned data sets, we will estimate noise on a much larger scale, using HTRC Extracted Features to associate each English-language volume in HathiTrust with an estimated level of transcription error. We will also map paratext, and identify (in volumes of fiction, poetry, and biography) pages that do not belong to the specified genre. (Running headers are already separated from body text in the extracted features dataset.)

Notebooks to help researchers estimate uncertainty. For each of the text analysis methods described above, we will produce a Jupyter notebook that helps researchers infer the level of uncertainty to be expected for a given level of error in their corpus. If researchers are working with HathiTrust volumes, they will be able to estimate error simply by passing in a list of volume IDs; if they are working with a different corpus, they can provide an estimate or a text sample. Since these notebooks will not be able to envision all possible second-order comparative effects, we will also provide

A report that summarizes conclusions and makes recommendations. Worked examples and case studies will demonstrate how to estimate uncertainty for downstream conclusions based on noisy text. The report will also, secondarily, make recommendations for further work, aiming especially to advise funders about best practices for mapping paratext.

Prototypes to support further research. The Ichneumon project, directed by David A. Smith, may solve part of the paratext problem automatically, by contrasting works that exist in multiple editions. But we hope to lay the groundwork for solving the problem also in a broader set of cases, by training a model that can remove paratext automatically from works of fiction and volumes of poetry. Implementing this model at library scale is beyond the scope of this project. But we will produce a prototype and assess its reliability.

e. Methodology and standards.

Since we will be aligning optically-transcribed text with manual ground truth, OCR accuracy can be measured directly: we will express it as the harmonic mean of word-level precision and recall (F1 score). We will also borrow a strategy cleverly deployed in recent work by Hill and Hengchen—assigning a level of error to each page in each book, so we can plot the deformation of results across levels of error.

While we use volume descriptors and MARC metadata drawn from HathiTrust, we will rely on the Internet Archive for the OCR in our 1,100-volume dataset, so it can be made freely available to researchers. We will draw metadata about book reviews from previous work by Underwood (both *Distant Horizons* and work in progress).

Jupyter notebooks will be written in Python 3. In developing a prototype to separate paratext from volumes of poetry and fiction, we will build on the strategy successfully adopted in McConnaughey et al.—a bidirectional sequence-labeling LSTM.

f. Work plan.

March-May 2020: Begin construction of the aligned corpus by selecting 1,100 texts available in manually and optically transcribed versions. Identify corrected Ichneumon texts where available.

June-July 2020: Align these texts at a page level. Train and test models that identify paratext (previous experience with this problem leads us to expect that performance will be best if three different models are trained for fiction, poetry, and nonfiction).

August-October 2020. Test different text analysis methods.

November-January 2020. Design Jupyter notebooks that help researchers estimate uncertainty.

February-May 2020. Estimate levels of error and paratextual noise across English-language volumes HathiTrust. Write the final project report.

g. Staff.

The PI, Ted Underwood, has extensive experience with the humanistic goals of text mining and with the technical issues that arise in OCR correction and paratext separation. He will define research questions that serve as case studies, co-author the final report, and train other staff involved in the project; he will also assist in corpus construction.

Wenyi Shang, a doctoral student in Information Sciences, will be the other primary contributor. Shang already has extensive experience in text mining, with publications that explore both Chinese history and English poetry. Shang will carry out many of the experiments, write Jupyter notebooks, and co-author the final report; he will also have primary responsibility for the Chinese-language corpus.

Other contributors to this cross-disciplinary team will include undergraduates and an English graduate student yet to be named.

h. Audience, evaluation, and sustainability.

The primary audiences for this project are the humanistic researchers who use text mining, and the librarians who develop and organize digital libraries. Both communities need a better understanding of the barriers to research using digital texts.

We will communicate and evaluate our findings using ordinary channels of publication and peer review. The final report for this project may be too long to submit *in toto*, but Underwood and Shang will write an article summarizing conclusions of central interest to researchers and submit this article to a widely-read journal such as *Digital Scholarship in the Humanities* or *The Journal of Cultural Analytics.* Libraries of aligned manually-transcribed and optically-transcribed texts, with and without paratext, can be associated with the article as

supporting data; supporting code will include Jupyter notebooks that guide researchers in estimating uncertainty.

Since the project is founded on a strong existing collaboration with HathiTrust, it will be easy to communicate our findings there. As explained in a supporting letter, a map of transcription error and paratextual noise in English-language books will be hosted at HathiTrust Research Center. To reach a digital library audience beyond HathiTrust, Underwood and/or Shang will also submit a paper to the Joint Conference on Digital Libraries.

This project has been designed to make sustainability simple. We are not building an interactive website; all deliverables are static documents, data sets, or code notebooks that can be hosted by existing institutions or journals, with copies of code in GitHub for easy access and Zenodo for permanence. The only challenge for sustainability is that HathiTrust will keep growing. Our map of error in the library will gradually date as new books are added. Pagination does also sometimes change in existing books. This is why we will pair the map of error to a specific release of HathiTrust Research Center Extracted Features. Extracted features provide a snapshot of the library that is guaranteed to remain stable.

When a new release of extracted features is generated—which happens on a three to four year schedule—the map could be generated anew. We will do that as needed. However, by four years out we hope to have an even better solution. What many researchers really want is not a map of noise in volumes (or items, to use FRBR terminology), but a library of relatively clean body texts that represent expressions or works. The methods being developed by Ichneumon make it possible to envision a library of that kind, or at least a set of tools that assist researchers in extracting work-level texts.

NATIONAL ENDOWMENT FOR THE Humanities

Budget Form

OMB No 3136-0134
Expires 6/30/2021

Applicant Institution: *University of Illinois*
Project Director: *William Underwood*
Project Grant Period: *3/01/2020-5/31/2021*

click for Budget Instructions

	Computational Details/Notes	(notes)	Year 1	(notes)	Year 2	(notes)		Project Total
			03/01/2020-02/28/2021__		03/01/2021-05/31/2021			
1. Salaries & Wages								
PD; William Underwood	12% one summer month in Year 1	12%	$1,497	0%	$0	%		$1,497
GRA	50% FTE for 11 months in Year One, 50% FTE for 2 months in Year 2	50%	$24,310	50%	$4,117	%		$28,427
Graduate hourly	100 hours; $18/hr in Year 1	%	$1,800	%		%		$1,800
		%		%		%		$0
		%		%		%		$0
		%		%		%		$0
2. Fringe Benefits								
Project Director	41.98%		$628					$628
GRA	8.02%		$1,950		$0			$1,950
Graduate hourly	0.10%		$2					$2
3. Consultant Fees								
								$0
4. Travel								
								$0
								$0

5. Supplies & Materials								
								$0
6. Subawards								
								$0
7. Other Costs								
Tuition Remission	64% of GRA monthly stipend		$15,558		$2,635			$18,193
8. Total Direct Costs	**Per Year**		**$45,745**		**$6,752**		**$0**	**$52,497**
9. Total Indirect Costs								
a. Rate: 58.6%								
b. Federal Agency: ONR	**Per Year**		$17,689		$2,606		$0	**$20,295**
Effective Period: 07/01/2019-06/30/2020								
10. Total Project Costs	(Direct and Indirect costs for entire project)							**$72,792**

11. Project Funding	**a. Requested from NEH**	Outright:	$73,122
		Federal Matching Funds:	$0
		TOTAL REQUESTED FROM NEH:	**$73,122**
	b. Cost Sharing	Applicant's Contributions:	$0
		Third-Party Cash Contributions:	
		hird-Party In-Kind Contributions:	$0
		Project Income:	$0
		Other Federal Agencies:	$0
		TOTAL COST SHARING:	**$0**
12. Total Project Funding			**$73,122**

Total Project Costs must be equal to Total Project Funding ----> ($72,792 = $73,122 ?)
Third-Party Contributions must be

greater than or equal to Requested Federal Matching Funds ----->	(	$0	≥	$0	?)

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
BUDGET JUSTIFICATION
Principal Investigator: William Underwood
Period: March 1, 2020 – May 31, 2021

Project Dollars

Salaries and Wages **$31,724**

Funding to support PI, William Underwood, 0.12 of one summer month in year one. Professor Underwood will oversee the scope of work of the proposed project. ($1,497)

One 50% iSchool GRA is requested for nine academic and two summer months in year one of the project, and for two months into the second year period. One 17% English GRA is requested for one semester, 4.5 months, in year one of the project. One undergraduate hourly is requested to work approximately 100 hours in year one, at $18/hour. ($30,227)

The iSchool GRA will have primary responsibility for many experiments in the project, and for building a Chinese-language corpus. The English GRA, and undergraduate hourly worker, will assist project staff with construction of an English-language corpus.

Salaries are based on actual UIUC AY2019 rates and are incremented at a rate of 3.0% each year.

Fringe Benefits **$2,910**

Fringe benefits are charged at a rate of 41.98% on faculty salaries. Benefits include retirement, worker's compensation, health, life and dental insurance, termination, and Medicare. Fringe benefits are charged at a rate of 8.02% on graduate student salaries. Benefits include worker's compensation and health, life and dental insurance. Fringe benefits are charged at a rate of .10%. Benefits include worker's compensation.

Consultant Fees **$0.0**

N/A

Travel **$0.0**

N/A

Supplies and Materials **$0.0**

N/A

Subawards **$0.0**

N/A

Other Costs **$18,193**

This budget category includes tuition remission assessed at 64.0% of graduate student salaries.

Total Direct Costs: $52,827

Indirect Costs **$20,295**

Indirect costs are assessed at a rate of 58.6% of Modified Total Direct Costs (MTDC). MTDC is direct costs less equipment, tuition remission, and subawards in excess of $25,000.

Total Project Costs: $73,122

5. Appendices.

a. Environmental scan.

Several useful studies have evaluated the effects of errorful OCR transcription on information retrieval (see Tanner et al.), and some have evaluated its effects on natural language processing (Lopresti). There are also important recent works that assess the effects of OCR error on text mining methods including, for instance, topic modeling and authorship attribution (Franzini et al.; Hill and Hengchen). However, these studies tend to focus on questions that have a relatively narrow chronological scope. For instance, authorship attribution is usually bounded by an active career that may only last two or three decades.

This is an awkward limitation, since contemporary humanists are increasingly using text mining to explore large comparative questions and long timelines (Armitage and Guldi; Underwood). When research questions span a range of social contexts, the absolute level of error in a corpus can become less important than its distribution across time or between genres and social strata. Few studies address this problem directly. Moreover, as we showed in Figure 1 above, paratext often creates more significant distortion for text mining than transcription error as such. But while there is a great deal of excellent work on methods for identifying paratext (McConnaughey et al.; Smith et al., "Detecting) there have been few (if any) studies of the effect of paratextual contamination on large-scale text mining. In practice, researchers have proceeded by trying to minimize distortion, without first fully mapping the potential for error.

b. Bibliography.

i. Related works by project staff.

Capitanu, Boris, Ted Underwood, Peter Organisciak, Timothy Cole, Maria Janina Sarol, J. Stephen Downie. *The HathiTrust Research Center Extracted Feature Dataset* (1.0) [Dataset]. HathiTrust Research Center, 2016. http://dx.doi.org/10.13012/J8X63JT3.

Underwood, Ted. *Distant Horizons: Digital Evidence and Literary Change.* Chicago: University of Chicago Press, 2019.

Underwood, Ted, Mike Black, Loretta Auvil, and Boris Capitanu. "Mapping Mutable Genres in Structurally Complex Volumes." 2013 IEEE International Conference on Big Data, pp. 95-104.

Underwood, Ted, Boris Capitanu, Peter Organisciak, Sayan Bhattacharyya, Loretta Auvil, Colleen Fallaw, J. Stephen Downie. *Word Frequencies in English-Language Literature, 1700-1922 (0.2) [Dataset].* HathiTrust Research Center, 2015. doi:10.13012/J8JW8BSJ. https://sharc.hathitrust.org/genre

Underwood, Ted. "Page-Level Genre Metadata for English-Language Volumes in HathiTrust, 1700-1922." figshare. 2014. Covers 854,476 volumes. http://dx.doi.org/10.6084/m9.figshare.1279201

ii. Other works referenced.

Armitage, David, and Jo Guldi. *The History Manifesto.* Cambridge: Cambridge University Press, 2014.

Eder, Maciej. "Mind your corpus: systematic errors in authorship attribution." *Literary and Linguistic Computing* 28 (2013): 603–14. doi:10.1093/llc/fqt039.

Franzini, Greta, Mike Kestemont, Gabriela Rotari, Melina Jander, Jeremi K. Ochab, Emily Franzini, Joanna Byszuk, and Jan Rybicki. "Attributing authorship in the noisy digitized correspondence of Jacob and Wilhelm Grimm." *Frontiers in Digital Humanities* 5, 2018. https:/doi.org/10.3389/fdigh.2018.00004.

Hill, Mark J., and Simon Hengchen, "Quantifying the Impact of Dirty OCR on Historical Text Analysis," *Digital Scholarship in the Humanities,* 22 April 2019, https://doi.org/10.1093/llc/fqz024.
Quinn, William. "Andrew W. Mellon Foundation Funds Northeastern's NULab to Study OCR for the Humanities." July 22, 2017. https://ocr.northeastern.edu/andrew-w-mellon-foundation-funds-northeasterns-nulab-to-study-ocr-for-the-humanities/

Lopresti, Daniel. "Optical Character Recognition Errors and their Effects on Natural Language Processing." *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data* (Singapore, 2008), pp. 9-18.

McConnaughey, Lara, Jennifer Dai, and David Bamman. "The Labeled Segmentation of Printed Books." *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* 2017, https://www.aclweb.org/anthology/D17-1077.

Smith, David A. "Tracking Reader Annotations in Printed Books by Collating and Transcribing Multiple Exemplars," NEH Digital Advancement Grant Proposal, 2018.

Smith, David A., and Ryan Cordell. "A Research Agenda for Historical and Multilingual Optical Character Recognition," NULab 2018, https://ocr.northeastern.edu/report/.

Smith, David A.. Ryan Cordell, Elizabeth Maddock Dillon, Nick Stramp, and John Wilkerson. Detecting and modeling local text reuse. In *Proceedings of the ACM+IEEE-CS Joint Conference on Digital Libraries (JCDL)*, 2014.

Tanner, Simon, Trevor Muñoz, and Pich Hemy Ros. “Measuring mass text digitization quality and usefulness: Lessons learned from assessing the OCR accuracy of the British Library’s 19th century online newspaper archive.” *D-Lib Magazine*, 15(7/8), 2009. URL http://www.dlib.org/dlib/july09/munoz/07munoz.html.

Ted Underwood

Professor, School of Information Sciences and Department of English
University of Illinois at Urbana-Champaign
tunder@illinois.edu

Background

Education

Ph.D., English, Cornell University, January 1997

B.A., Philosophy, Williams College, 1989

Employment

Professor of Information Sciences, UIUC (2016 -)

Professor of English, UIUC (2014-)

Associate Professor of English, UIUC (2007-14)

Assistant Professor of English, UIUC (2003-06)

Assistant Professor of English, Colby College (1998-2003)

Visiting Assistant Professor, University of Rochester (1997-8)

Publications

Works in progress

- *A Perspectival History of Fiction in English.*
- "Reviews of English-Language Fiction" (with Kent Chang, Yuerong Hu, and Jessica Witte).

Books

- *Distant Horizons: Digital Evidence and Literary Change.* Chicago: Univ. of Chicago Press, 2019.
- *Why Literary Periods Mattered: Historical Contrast and the Prestige of English Studies.* Palo Alto, CA: Stanford University Press, 2013.
- *The Work of the Sun: Literature, Science, and Political Economy 1760-1860*. New York: Palgrave, 2005.

Recent papers in proceedings

- "The Historical Significance of Textual Distances," 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature.
- David Bamman, Ted Underwood, Noah Smith. "A Bayesian Mixed Effects Model of Literary Character." Association for Computational Linguistics 2014.
- Ted Underwood, Mike Black, Loretta Auvil, Boris Capitanu. "Mapping Mutable Genres in Structurally Complex Volumes." 2013 IEEE International Conference on Big Data, pp. 95-104.

Recent journal articles and chapters

- "Why Literary Time is Measured in Minutes." *ELH* 85.2 (2018): 341-65.
- "Algorithmic Modeling; or, Modeling Data We Do Not Yet Understand." In *The Shape of Data in Digital Humanities,* ed. Julia Flanders and Fotis Jannidis. Ashgate, 2018.
- Ted Underwood, David Bamman, and Sabrina Lee. "The Transformation of Gender in English-Language Fiction." *Journal of Cultural Analytics,* February 2018.
- "A Genealogy of Distant Reading." *Digital Humanities Quarterly* 11.2 (2017).

- James F. English and Ted Underwood. "Shifting Scales: Between Literature and Social Science," introduction to a special issue co-edited by English and Underwood, *Modern Language Quarterly* 77.3 (2016): 277-95.
- Ted Underwood and Jordan Sellers. "The Longue Durée of Literary Prestige." *Modern Language Quarterly* 77.3 (2016): 321-44.
- "The Life Cycles of Genres." *Journal of Cultural Analytics,* May 2016.
- "Distant Reading and Recent Intellectual History." Debates in Digital Humanities 2016, ed. Matthew K. Gold and Lauren Klein. Minneapolis: University of Minnesota Press, 530-33.
- "Hold On Loosely: Or, Gesellschaft and Gemeinschaft on the Web," Debates in Digital Humanities 2016, ed. Matthew K. Gold and Lauren Klein. Minneapolis: University of Minnesota Press, 519-22.
- "The Literary Uses of High-Dimensional Space." In Assumptions of Sociality: A Colloquium of Social and Cultural Scientists, a special issue of *Big Data and Society* ed. John Mohr, Ronald Breiger and Robin Wagner-Pacifici. 2015.
- Andrew Goldstone and Ted Underwood. "The Quiet Transformations of Literary Study: What Thirteen Thousand Scholars Could Tell Us." *New Literary History* 45.3 (2014): 359-84.

Software and data

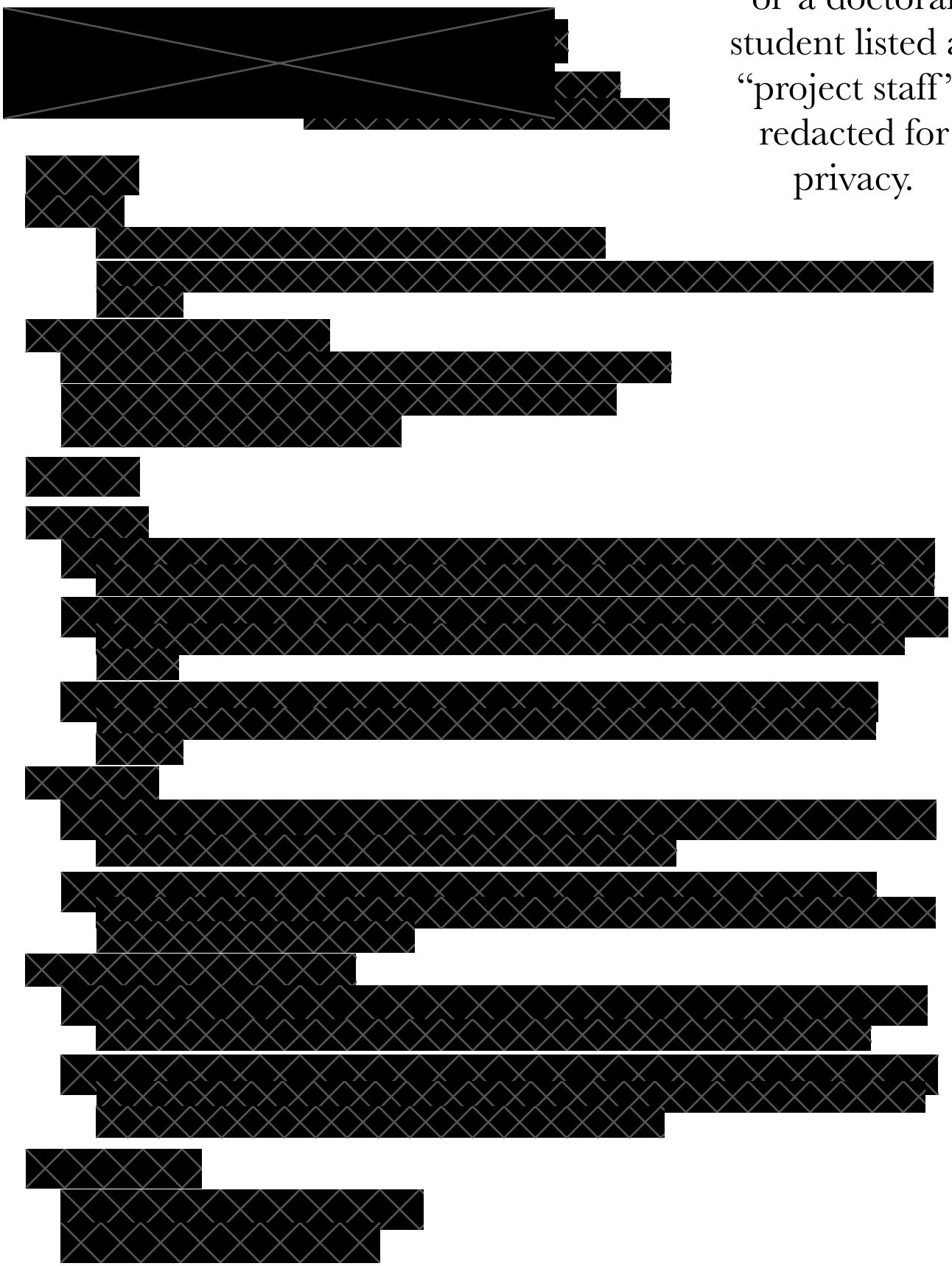
- Ted Underwood, Boris Capitanu, Peter Organisciak, Sayan Bhattacharyya, Loretta Auvil, Colleen Fallaw, J. Stephen Downie (2015). *Word Frequencies in English-Language Literature, 1700-1922 (0.2) [Dataset]*. HathiTrust Research Center. doi:10.13012/J8JW8BSJ. https://sharc.hathitrust.org/genre
- Ted Underwood, "Page-Level Genre Metadata for English-Language Volumes in HathiTrust, 1700-1922." figshare. 2014. Covers 854,476 volumes. http://dx.doi.org/10.6084/m9.figshare.1279201

Grants and honors

- Fellow, National Humanities Center, Research Triangle, NC, 2018-19, $51,000.
- Collaborator, WCSA+DC (J. Stephen Downie and Beth Plale, co-PIs). Andrew W. Mellon Foundation, $1.17 million.
- Associate, Center for Advanced Study, UIUC 2015-16.
- Primary collaborator, Text Mining the Novel (Andrew Piper, McGill, PI). My portion of the project funded to $120,000 CDN over six years (2014-20).
- ACLS Digital Innovation Fellowship, calendar year 2014. $85,000.
- LAS Centennial Scholar, University of Illinois, Urbana-Champaign, 2013-16. $30,000.
- NEH Digital Humanities Start-Up Grant, 2013-14. $57,100.

Coverage

- Dan Sinykin, "How Computational Analysis is Teaching Us to Read in New Ways," *The Washington Post,* July 30, 2018, https://www.washingtonpost.com/news/posteverything/wp/2018/07/30/how-computational-analysis-is-teaching-us-to-read-in-new-ways/?utm_term=.116777a24e13
- "Machines are Getting Better at Literary Analysis," *The Economist,* March 8, 2018, https://www.economist.com/prospero/2018/03/08/machines-are-getting-better-at-literary-analysis
- Kat Eschner, "Women were Better Represented in Victorian Novels than Modern Ones," *Smithsonian Magazine,* February 14, 2018, https://www.smithsonianmag.com/arts-culture/what-big-data-can-tell-us-about-women-and-novels-180968153/



This was the vita of a doctoral student listed as "project staff"; redacted for privacy.

6. Project participants, consultants, and advisors.

Downie, J. Stephen. Professor, Associate Dean for Research, and co-director of HathiTrust Research Center, School of Information Sciences, University of Illinois, Urbana-Champaign.

Shang, Wenyi. Doctoral student in the School of Information Sciences, University of Illinois, Urbana-Champaign.

Smith, David A. Associate Professor, Khoury College of Computer Sciences, Northeastern University.

Underwood, Ted. PI, Professor of Information Sciences and English, University of Illinois, Urbana-Champaign.

Northeastern University
Khoury College of Computer Sciences

Professor Ted Underwood
School of Information Sciences and Department of English
University of Illinois at Urbana-Champaign

10 May 2019

Dear Prof. Underwood,

I am writing to confirm my commitment to provide machine learning models and data for text collation and OCR correction for your project, "Broadening Access to Text Analysis by Describing Uncertainty."

Khoury College of Computer Sciences
202 West Village H
360 Huntington Avenue
Boston, MA 02115
617.373.2462
www.ccs.neu.edu/home/dasmith
dasmith@ccs.neu.edu

I have read your proposal with great interest and believe the project will provide useful guidance for research in literary studies and natural language processing. Practitioners in both fields in will benefit from new methods understanding and communicating the effects of inference on mass-digitized repositories with uneven transcription accuracy and structural metadata.

Our team at Northeastern University will be happy to provide assistance applying our open-source software for text alignment, collation, and correction to the scanned English and Chinese books in your reference corpus. We will also compare these texts to data in larger corpora from the Internet Archive and HathiTrust, which can provide additional evidence for text transcription and structure.

I am excited about the prospects for this project and wish you success!

Sincerely,

David A. Smith
Associate Professor

ILLINOIS

SCHOOL OF INFORMATION SCIENCES
501 E. Daniel St., MC-493
Champaign, IL 61820-6211

12 May 2019

To Whom It May Concern:

I am writing as the Illinois co-director of HathiTrust Research Center (HTRC), in order to express our commitment to the grant proposal 'Broadening Access to Text Analysis by Describing Uncertainty' (Ted Underwood, PI). As co-director of HTRC, I have collaborated with Underwood frequently. He attends most of our staff meetings and serves as a senior advisor to HTRC, helping us especially by modeling the needs of our users interested in text-mining. In particular, Underwood has collaborated with the subcommittee designing Extracted Features, for instance by helping our senior research programmer, Boris Capitanu, scope out the rationale for separating recurring headers and footers from paratext.

Dr. Underwood is one of the leading researchers in the emerging field of cultural analytics. My experience collaborating with him gives me confidence that he is the right person to guide research on the challenges that error and noise pose for researchers doing text mining in large digital libraries.

HTRC is committed to supporting this research by providing a data capsule to support text mining on books in copyright, in both English and Chinese. We will also help publicize the results of the research, for instance by guiding researchers who use our Extracted Features to the proposed map of error and paratextual noise in English-language books, which will be keyed to pagination in our Extracted Feature data set. We see this as a solution to a recurring problem for our users.

I am standing by should you have any questions.

Sincerely,

J. Stephen Downie, PhD

Associate Dean for Research & Professor
Co-director, HathiTrust Research Center

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
217.333.3280 • ischool@illinois.edu • ischool.illinois.edu

8. Dissemination plan.

The project has two primary audiences: the humanists who use text mining in their research, and the librarians who develop and organize digital libraries.

To reach the first audience, Underwood and Shang will write an article summarizing conclusions of central interest to researchers. This article will describe, for instance,

- Typical levels of uncertainty affecting frequently-used text analysis methods. (It might be dramatically effective for Underwood to revisit and revise one of his own previous conclusions.)

- The robustness of new neural methods in the presence of OCR error. Many scholars would like to use these methods, but their performance in noisy collections of historical text is not well understood.

- New methods for estimating uncertainty.

This article will be submitted to a widely-read journal such as *Digital Scholarship in the Humanities* or *The Journal of Cultural Analytics.* Libraries of aligned manually-transcribed and optically-transcribed texts, with and without paratext, can be associated with the article as supporting data; supporting code will include Jupyter notebooks that guide researchers in estimating uncertainty.

Since the project is founded on a strong existing collaboration with HathiTrust, it will be easy to communicate our findings there. As explained in a supporting letter, a map of transcription error and paratextual noise in English-language books will be hosted at HathiTrust Research Center.

To reach a digital library audience beyond HathiTrust, Underwood and/or Shang will also submit a paper to the Joint Conference on Digital Libraries. This paper will emphasize conclusions of interest to librarians and information scientists—for instance, new methods for identifying and mapping paratext.

DEPARTMENT OF THE NAVY
OFFICE OF NAVAL RESEARCH
875 NORTH RANDOLPH STREET
SUITE 1425
ARLINGTON, VA 22203-1995

Agreement Date: December 21, 2017

NEGOTIATION AGREEMENT

INSTITUTION: **THE UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN**
CHAMPAIGN, ILLINOIS 61820-6242

The Facilities and Administrative (F&A) rates contained herein are for use on grants, contracts and/or other agreements issued or awarded to The University Of Illinois at Urbana-Champaign by all Federal Agencies of the United States of America, in accordance with the cost principles mandated by 2 CFR Part 200. These rates shall be used for forward pricing and billing purposes for the University of Illinois at Urbana Champaign Fiscal Year 2020. This rate agreement supersedes all previous rate agreements/determinations for Fiscal Year 2020.

Section I: RATES - TYPE: PREDETERMINED (PRED)

Type	From	To	On-Campus Rates	Off-Campus Rates	Base	Applicable Function
PRED	7/1/19	6/30/20	58.6%	26.0%	(a)	Organized Research
PRED	7/1/19	6/30/20	45.8%	26.0%	(a)	Sponsored Instruction
PRED	7/1/19	6/30/20	31.9%	23.6%	(a)	Other Sponsored Activities

DISTRIBUTION BASE:

(a) Modified Total Direct Cost (MTDC), as defined in 2 CFR Part 200, consisting of all salaries and wages, fringe benefits, materials and supplies, services, travel, and subawards up to the first $25,000 each subaward (regardless of the period covered by the subaward); and excluding equipment (defined in Section II, paragraph G.1.), capital expenditures, charges for patient care and tuition remission, rental costs, scholarships and fellowships, participant support costs as well as the portion of each subaward in excess of $25,000.

SECTION II: GENERAL TERMS AND CONDITIONS

A. **LIMITATIONS:** Use of the rates set forth under Section I is subject to any statutory or administrative limitations and is applicable to a given grant, contract or other agreement only to the extent that funds are available and consistent with any and all limitations of cost clauses or provisions, if any, contained therein. Acceptance of any or all of the rates agreed to herein is predicated upon all the following conditions: (1) that no costs other than those incurred by the institution were included in its indirect cost pool as finally accepted and that all such costs are legal obligations of the institution and allowable under governing cost principles; (2) that the same costs that have been treated as indirect costs are not claimed as direct costs; (3) that similar types of costs, in like circumstances, have been

accorded consistent accounting treatment; (4) that the information provided by the institution, which was used as the basis for the acceptance of the rates agreed to herein and expressly relied upon by the Government in negotiating the said rates, is not subsequently found to be materially incomplete or inaccurate.

B. **ACCOUNTING CHANGES**: The rates contained in Section I of this agreement are based on the accounting system in effect at the time this agreement was negotiated. Changes to the method(s) of accounting for costs, which affects the amount of reimbursement resulting from the use of these rates, require the prior written approval of the authorized representative of the cognizant negotian agency for the Government prior to implementation of any such changes. Such changes include but are not limited to changes in the charging of a particular type of cost from indirect to direct. Failure to obtain such approval may result in subsequent cost disallowances.

C. **PREDETERMINED RATES**: The predetermined rates contained in this agreement are not subject to adjustment in accordance with the provisions of 2 CFR Part 200, subject to the limitations contained in Part A of this section.

D. **USE BY OTHER FEDERAL AGENCIES**: The rates set forth in Section I hereof are negotiated in accordance with and under the authority set forth in 2 CFR Part 200. Accordingly, such rates shall be applied to the extent provided in such regulations to grants, contracts and other agreements to which 2 CFR Part 200 applies, subject to any limitations in part A of this section. Copies of this document may be provided by either party to other Federal agencies to provide such agencies with documentary notice of this agreement and its terms and conditions.

E. **DFARS WAIVER**: Signature of this agreement by the authorized representative of the University of Illinois at Urbana-Champaign and the Government acknowledges and affirms the University's request to waive the prohibition contained in DFARS 231.303(1) and the Government's exercise of its discretion contained in DFARS 231.303(2) to waive the prohibition in DFARS 231.303(1). The waiver request by the University of Illinois at Urbana-Champaign is made to simplify the University's overall management of DOD cost reimbursements under DOD contracts.

F. **APPLICATION OF RATES:**

1. The rates included in Section I are not applicable to Intergovernmental Personnel Act (IPA) costs. If the University elects to seek reimbursement of F&A costs associated with IPA agreements, then the University and the Office of Naval Research shall establish a special indirect cost rate for IPA agreements in accordance with the provisions of 2 CFR Part 200.

2. Application of the appropriate On-Campus or Off-Campus indirect rate(s) is to be determined at the beginning of each sponsored agreement and is to be equitably adjusted if the circumstances which determined the application change materially during the period of performance.

a. The On-Campus rate is to be assessed except when a portion of the sponsored agreement is performed at an off-campus site. The criteria for utilization of the off-campus rate consists of all of the following: (a) performance at the off-campus site must be on a continuous basis; intermittent performance is not sufficient; (b) the University personnel working or engaged on the project must be physically located at an off-campus site; and (c) the off-campus

performance must be of sufficient duration; normally a full semester, summer term or the period of performance of the sponsored agreement. The off-campus rate will be used for the off-campus portion of the work on a sponsored agreement.

b. Off-campus costs may include costs incurred at the off-campus site for salaries, related benefits, supplies, utility costs, rent, local travel and other similar costs, which are treated as direct. Travel to and from an off-campus site is considered an off-campus cost.

G. **SPECIAL REMARKS:**

1. Equipment is defined as follows: The costs of items of purchased equipment with an estimated life of more than one year and an acquisition cost of $5,000.00 or more per unit, including lease-purchase agreements. This definition includes the cost of component parts, materials and/or supplies used to fabricate an item or piece of equipment when (1) the aggregate cost of the component parts, materials and/or supplies is $5,000.00 or more in value and (2) the cost of fabrication is documented and accounted for by the department.

2. The Government's agreement with the rates set forth in Section I is not an acceptance of the University of Illinois at Urbana-Champaign's accounting practices or methodologies. Any reliance by the Government on cost data or methodologies submitted by the University of Illinois at Urbana-Champaign is on a non-precedence-setting basis and does not imply Government acceptance.

3. In accordance with 2 CFR 200.414(g), the University of Illinois at Urbana-Champaign has requested an extension of its Fiscal Year FY 2019 rates. Therefore, the rates identified in Section I are an extension of the FY 2019 rates.

FOR THE UNIVERSITY:

Avijit Ghosh
VP Chief Financial Officer &
Authorized Representative of the
Board of Trustees of the University of Illinois

12/21/2017
Date

FOR THE GOVERNMENT:

SNYDER.BETH.A.1379326664

Digitally signed by SNYDER.BETH.A.1379326664
DN: c=US, o=U.S. Government, ou=DoD, ou=PKI, ou=USN, cn=SNYDER.BETH.A.1379326664
Date: 2017.12.22 14:45:44 -05'00'

Beth A. Snyder
Contracting Officer

Date

For information concerning this agreement contact:
Beth Snyder
Office of Naval Research
Phone: (703) 696-5755
E-mail: beth.snyder@navy.mil

INDIVIDUAL RATE COMPONENTS

Institution: University of Illinois at Urbana-Champaign

Type of Rate: Predetermined

FY Covered by Rate: 2020
(Extension of FY 2019 Rates)

	Organized Research FY 2020		Instruction FY 2020		Other Sponsored Activities FY 2020	
Negotiation Base: MTDC ($000's)	$248,397	$8,414	$569,496	$1,175	$64,373	$14,194
RATE COMPONENTS:	On Campus	Off Campus	On Campus	Off Campus	On Campus	Off Campus
1. ADMINISTRATIVE						
A. GA&GE	6.58%	6.58%	6.90%	6.90%	6.61%	6.61%
B. DA	10.77%	10.77%	14.08%	14.08%	7.71%	7.71%
C. DA Allowance	3.60%	3.60%	3.60%	3.60%	3.60%	3.60%
D. SPA	5.59%	5.59%	4.74%	4.74%	5.68%	5.68%
E. Student Admin. & Services	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
F. Adjustment for 26% Cap	(0.54%)	(0.54%)	(3.32%)	(3.32%)		
Subtotal	26.00%	26.00%	26.00%	26.00%	23.60%	23.60%
2. DEPRECIATION:						
A. Buildings & Improvements	4.31%		1.51%		1.17%	
B. Equipment	2.32%		0.49%		0.31%	
3. Interest	1.61%		0.51%		0.31%	
4. O&M	19.25%		8.23%		6.47%	
5. Library	3.81%		9.05%		0.00%	
Subtotal	31.32%		19.79%		8.26%	
6. Utility Cost Adjustment	1.30%					
TOTAL	**58.6%**	**26.0%**	**45.8%**	**26.0%**	**31.9%**	**23.6%**

FOR THE INSTITUTION:

Avijit Ghosh
VP Chief Financial Officer &
Authorized Representative of the Board of Trustees
of the University of Illinois

12/21/2017
Date

FOR THE GOVERNMENT:

Beth A. Snyder
Contracting Officer

Date

For Official Use Only

INDIVIDUAL RATE COMPONENTS