ACLS
Advancing the Humanities

ACLS Digital Innovation Fellowship

Deadline: 10/2/2012

CONTACT INFORMATION

Full Name, with salutation	**Professor Ted Underwood**		
Primary Email Address	**tunder@illinois.edu**		
Office Address	**208 English Building** **608 S. Wright St**		
City	**Urbana**		
State	**Illinois**	Zip	**61801**
Telephone	**217-351-5122**	Fax	
Country *(if NOT U.S.)*		Phone *(if NOT U.S.)*	
Home Address	**2411 Branch Rd**		
City	**Champaign**		
State	**Illinois**	Zip	**61822**
Telephone	**217-351-5122**	Fax	
Country *(if NOT U.S.)*		Phone *(if NOT U.S.)*	

Which is your preferred mailing address? **Home**

*(**Be sure to indicate your preferred mailing address.** This is where we will mail [regular first-class US postal service] the letter informing you of the result of your application.)*

EDUCATION

PhD received from	**Cornell University**
PhD received	**5/1997**
PhD major discipline	**English Language and Literature**
Title of doctoral dissertation	**"The Labor Done by the Sun: Science, Poetry and Political Economy 1760-1860"**
Name of dissertation supervisor	**Reeve Parker**
Master's degree received from	
Degree	
Date master's degree received	
Master's degree major discipline	
BA/BS received from	**Williams College**
Date BA/BS received	**5/1989**
BA/BS major discipline	**Philosophy**

List any additional degrees

List up to six foreign languages you can use, indicating proficiency in reading, speaking, and writing. *(Use E=Excellent, G=Good, F=Fair or less, NA=Not applicable.) If you are either a Native Speaker or Heritage Speaker of a language, please indicate by checking the appropriate box.*

				Native	*Heritage*
Language	*Reading*	*Speaking*	*Writing*	*Speaker*	*Speaker*
French	**G**	**F**	**F**		
German	**F**	**F**	**F**		

CURRENT POSITION

Rank/Title	**Associate Professor**
Discipline	**English**
Specialization	**digital humanities**
Department	**English**
Institution	**University of Illinois, Urbana-Champaign**
Date you began this position	**8/2003**
Are you tenured?	**Y**
If yes, when did your first tenured semester begin?	**8/2008**
If you are an Assistant Professor or equivalent, when did you begin your first teaching semester/quarter at that rank, even if that occurred in a previous job?	
If you do NOT hold the rank of Full, Associate, or Assistant Professor, as a research scholar, with which group would you most identify?	
Second Institution *(if applicable)*	
Date you began this position	
If you do not hold an academic job, what is your current position?	

PROFESSIONAL BACKGROUND

List positions held (professional, teaching, administrative, curatorial) since college graduation, beginning with current position. Give name of institution, title, and approximate dates of employment for each.

Institution/Employer **University of Illinois, Urbana-Champaign**

Title **Associate Professor**

From **8/2007** To **10/2012**

Institution/Employer **University of Illinois, Urbana-Champaign**

Title **Assistant Professor**

From **8/2003** To **5/2007**

Institution/Employer **Colby College**

Title **Assistant Professor**

From **8/1998** To **5/2003**

Institution/Employer **University of Rochester**

Title **Visiting Assistant Professor**

From **8/1997** To **5/1998**

Institution/Employer

Title

From To

Please provide any other relevant information to help reviewers better understand your professional background and to contextualize elements of your career listed elsewhere in your application. Some possible issues include service, teaching, administration, family and other personal circumstances, public humanities work, alternate career paths, character of your work (archival, field work, collaborative, etc.).

In the 1980s, I worked as a software developer for Artificial Intelligence Atlanta, developing a prototype of a file retrieval expert system. That work was in Prolog; I'm currently fluent in Java, R, Python, and various relational databases. I've also immersed myself in the literature of machine learning. I don't try to produce original research in that discipline, but I do stay current in it.

AWARDS

Beginning with the most recent, list up to eight of the grants, fellowships, scholarships, academic honors, or awards you have received, giving in each case the dates, purposes (tuition, travel, expenses, etc.), and, if funded, the approximate amounts. If you are listing only selected awards, choose those that are most significant. Please do not be concerned if you cannot recall exact dates or amounts, and do not feel you must use all eight entries.

Award **Andrew W. Mellon Foundation, "Uses of Scale"**

Award Type **Non-ACLS Fellowship or Grant**

From **8/2012** To **8/2013** Amount **$40,000**

Purpose **Digital collaboration between three institutions.**

Award **Andrew W. Mellon Foundation, "Expanding SEASR"**

Award Type **Non-ACLS Fellowship or Grant**

From **8/2010** To **12/2012** Amount **$70,000**

Purpose **Paid RAs to help develop tools for text-mining.**

Award **William and Flora Hewlett Int'l Travel Grant**

Award Type **Non-ACLS Fellowship or Grant**

From **5/2005** To **7/2005** Amount **$3,000**

Purpose **Travel for book research.**

Award **Research Board Grant, UIUC**

Award Type **Non-ACLS Fellowship or Grant**

From **8/2004** To **5/2005** Amount

Purpose **Book completion fellowship.**

Award **Messenger-Chalmers Graduate Prize**

Award Type **Dissertation Award**

From **5/1997** To **5/1997** Amount

Purpose **Dissertation recognized for its historical rsch.**

Award **Harry Falkenau Service Award, Cornell University**

Award Type **Other**

From **5/1995** To **5/1995** Amount

Purpose **Recognize service to department.**

RESEARCH PROJECT

Research Proposal Title

Understanding Genre in a Collection of a Million Volumes

Research Proposal Abstract

Information about genre makes large digital collections much more useful, but is largely missing in our metadata. It is possible to recognize genre algorithmically, and a digital approach has several important advantages: it allows classification schemes to be modified at will, and allows membership in a genre to be a matter of degree rather than a hard boundary. I propose to develop software that can classify HathiTrust collections by genre. I will draw on existing machine learning research, but also modify it to fit this domain: for instance, genre classifications need to change continuously across the timeline, and generically heterogenous volumes need to be divided into parts. I will make my software available for other scholars, and use it to develop a book on 19c literary history.

If there is a web page associated with your project, please provide the URL here:

http://tedunderwood.wordpress.com/

Broad Humanistic Significance of Project:

Your proposal will be reviewed by scholars within your specific discipline and in other disciplines in the humanities and related social sciences. State the significance of your project for the humanities and related social sciences. Indicate how and why the project might be of interest to scholars in other disciplines. Please avoid discipline-specific jargon that may pose a problem for non-specialists.

This project will primarily be useful for literary scholars, with some relevance also for historians. Within literary studies, its utility is not limited by period or language. The test case for my methods will be a collection of a million English-language volumes, 1800-1949. But I'm working with HathiTrust Research Center, so my software can be embedded in their API, allowing other scholars to classify English-language works in other periods, or according to other classification schemes. Extending that service to other languages would require some additional work, but is not in principle a difficult problem.

But why would literary scholars want to classify large collections by genre? Isn't the concept of genre too fuzzy for computers to be useful? There are several different answers here, because genre itself is complex. Where the broadest categories are concerned (prose/verse, fiction/nonfiction), generic boundaries are fairly stable, and fairly easy to discern algorithmically. Metadata of this kind would be valuable for a broad range of projects, including historical projects that are not primarily concerned with genre. Even historians need to know whether the phenomena they're tracing appear in fiction or nonfiction works, and subject classifications (PQ, PR, PS, etc.) do not actually answer that question.

Where smaller, more volatile categories are concerned (poetic drama or the gothic novel), the boundaries of genres tend to overlap, and tend to be redrawn frequently by critics. But I would argue that these complexities actually strengthen the case for a digital approach to genre. Unlike crowdsourcing, algorithmic classification is easy to repeat; we can classify the same collection in dozens of different ways. Moreover, we can allow volumes to belong to multiple subgenres at once, assigning a different level of confidence to each tag. Treating membership in a genre as a question of degree can help us understand how genres emerge and change.

If you are planning to conduct your proposed research project in a particular location, please specify where and when you plan to do so.

First semester / Urbana-Champaign, second semester / Urbana-Champaign.

List any countries or geographical areas on which your research is focused.

1. **United Kingdom**

2. **United States**

3.

4.

Other

List any countries or geographical areas other than the US in which you have done research in the last five years.

1. **United Kingdom**

2.

3.

Other

Please identify up to five disciplinary areas, in order of relevance, that best describe your research project.

1.	**Digital Humanities**	Other
2.	**Languages and Literature**	Other
3.	**Information Studies**	Other
4.		Other
5.		Other

Name: Professor Ted Underwood

ADMINISTRATIVE INFORMATION

This information is REQUIRED (except as noted). It is for administrative purposes only and will not be distributed as part of the selection process.

Current salary *(do not add benefits or summer salary)* **$79,271**

Amount requested for STIPEND **$60,000**

Amount requested from ACLS for PROJECT COSTS **$25,000**

What is your country of citizenship? **United States**

If NOT United States, do you hold US Permanent Resident status?

AND have you lived in the US continuously for at least the past 3 years?

Beginning date for ACLS Digital Innovation Fellowship **8/27/2013**

Ending date for ACLS Digital Innovation Fellowship **8/27/2014**

If the ACLS Digital Innovation Fellowship tenure period and stipend requested will be used toward a longer research leave, please give dates of the total planned leave.

From To

List other sources of support, for example, sabbatical salary, other fellowships and grants, ALREADY CONFIRMED in connection with your proposed research project or planned total period of research leave. Also indicate the approximate amount of funding and period of support.

Source

From To Amount

Source

From To Amount

Source

From To Amount

List other major funding sources, with approximate amount and tenure period, to which you ARE APPLYING for your present research proposal.

Source **National Endowment for the Humanities**

From **5/2013** To **11/2014** Amount **$57,163**

Source **UIUC Research Board**

From **8/2013** To **12/2013** Amount **$25,000**

Source

From To Amount

The following questions are optional and will be used for statistical purposes only.

Date of birth **8/16/1968** Gender **M**

Name: Professor Ted Underwood

With which group(s) do you most identify?	**Y**	White (not of Hispanic origin)
		Black (not of Hispanic origin)
		Hispanic or Latino
		American Indian or Alaskan Native
		Asian
		Native Hawaiian or other Pacific Islander
		Other

The following questions are for informational purposes only.

1. How did you learn about ACLS fellowship programs?

Other/informal communication; Other: Twitter

2. Please identify the ACLS member scholarly societies or ACLS affiliate organizations (if any) of which you are a member or with which you have an affiliation.

Modern Language Association of America

3. Please identify all ACLS fellowship programs (if any) to which you have previously applied.

ACLS Fellowship; Ryskamp Fellowship

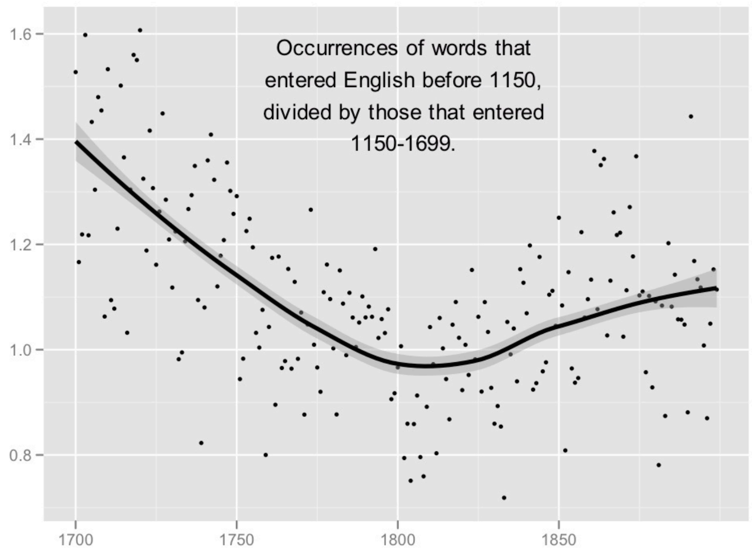
Understanding Genre in a Collection of a Million Volumes.

Since the 1990s, literary scholars have hoped that digital tools would permit them to explore the printed record on a fundamentally different scale — not the hundreds of books a single scholar can read, but the hundreds of thousands that have been preserved in modern periods (Olsen 1993). In some ways that promise is being fulfilled, but the scarcity of useful metadata still significantly limits our ability to interpret large collections.

It is difficult to make interesting arguments about the printed record, after all, if scholars have to treat it as a single pool of documents, differentiated only by publication date. That was the fundamental problem with Google's ngram viewer, and many digital projects are shaped by a need to detour around the same obstacle. We have seen a surge of interest in topic modeling over the last twelve months, for instance, because topic modeling is one of a few techniques that do allow researchers to extract patterns from undifferentiated documents. The results can be useful, especially for exploratory purposes. But to build literary arguments on large digital collections, we are usually going to need metadata that allow us to divide works into categories. Whether the printed record is divided geographically, demographically, or by subject, form, and genre, categorization creates the points of contrast that articulate a meaningful argument.

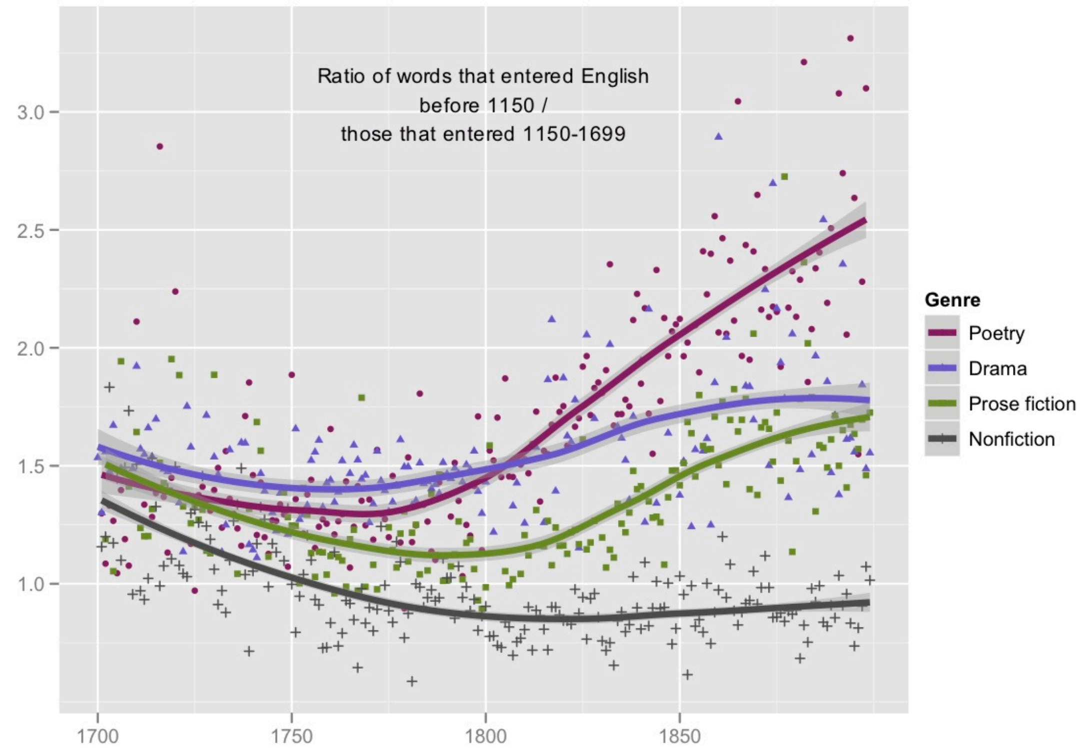
For literary scholars, genre is particularly valuable information, and categorization by genre can transform abstract questions into leads of significance for literary history. Let me illustrate with a brief example from my own work. I started with a hunch that English prose was more self-consciously learned around 1800 than it had been either earlier or later. To test that hunch, I borrowed linguists' observation that different parts of the English lexicon predominate in different "registers," or social contexts. Old English words are used more commonly in speech; words borrowed from French and Latin are used more often in writing, and especially formal writing (Bar-Ilan and Berman 2007). I developed a collection of 4,275 eighteenth- and nineteenth-century works (in collaboration with Jordan Sellers, TCP-ECCO and the Brown Women Writers

Project), and graphed the aggregate yearly ratio between pre- and post-twelfth-century parts of the lexicon. I excluded stopwords and words coined after 1699. The result confirmed my initial hypothesis, but was not otherwise terribly illuminating:



There is a shift toward more learned diction in the eighteenth century that reverses itself in the nineteenth. But since the older part of the lexicon is used more often in speech, the reason for the change might have been, for instance, that the collection contained less drama around 1800. To rule out that sort of explanation, I had to divide the collection by genre. So Jordan Sellers and I classified 4,275 works by hand. The results this time were quite startling.

In attempting to measure diachronic changes in style, we had stumbled on evidence of a differentiating process that cast new light on the definition of literature itself. It is already recognized that modern definitions of literature took shape between 1700 and 1900. In the early eighteenth century “literature” meant learning or written discourse generally; by the end of the nineteenth century the word described a category of writing set apart by specifically fictive and aesthetic aims. It now appears that this conceptual differentiation of literary and nonliterary genres was paralleled by a corresponding change in writing practices that set fiction, drama, and poetry apart from nonfiction prose on the level of diction.



In the nineteenth century, we discovered, the older part of the lexicon became two or three times more common in literary genres than in nonfiction — a differentiation that had not existed in the early eighteenth century. These findings could enrich English literary history in many ways. For instance, they correct prevailing histories of "poetic diction," which still tend to repeat William Wordsworth's claim to have reduced "the space of separation betwixt Prose and Metrical composition" (Wordsworth 1800). Poetry did lose some specialis'd orthography in the nineteenth century, but in a broader sense the "space of separation" between poetry and prose increased dramatically. In fact, since the same words tended to become more common in poetry, fiction, and drama, one could even say that "poetic diction" gave way in the nineteenth century to a generalized "literary diction."

These are the kinds of results that digital humanists have long been promising literary scholars: new leads of broad significance on topics that scholars have been debating for centuries. (For more on the thematic and social changes associated with the emergence of literary diction, see

Underwood and Sellers 2012.) In my case, I found that the breakthrough did not depend on developing a more sophicated data-mining algorithm, or marking up individual texts. It depended simply on enriching a collection of appropriate scale with appropriate kinds of metadata.

As one goal of my fellowship, I propose to develop the argument sketched above into a book that more thoroughly examines the linguistic differentiation of literary and nonliterary genres between 1700 and 1949. To do that, I'll need a generically classified collection much larger than 4,275 volumes. Obtaining access to the works is not a problem: I have already obtained 800,000 public-domain English-language volumes from HathiTrust, and I can work with HathiTrust Research Center to study works still under copyright.

However, manually classifying a million volumes is not easy — which leads me to the second and more important goal of this fellowship: developing software to automatically classify volumes by genre in an appropriately flexible way, to support not only my own research but the endeavors of other scholars. This is a challenging task for many reasons. While the Library of Congress does define genre/form headings, they are often missing in MARC records, and call numbers tell us about subject classifications rather than genres. Moreover, different scholars may not even want to use the same set of genre categories. Genres are not, after all, fixed and absolute entities. Scholars define new genres in order to pose particular research questions, and other scholars will redraw their boundaries to advance different inquiries.

That instability actually strengthens the argument for an algorithmic approach to genre. We might convince scores of human readers to classify a collection of a million volumes — once. But we clearly cannot crowdsource that work every time a scholar wants to redraw the boundary between science fiction and fantasy. Instead, we need a flexible, rapid way of dividing large collections into a given set of categories. Doing that would both allow us to explore the history of genre itself, and provide a foundation for other digital projects that become more meaningful when paired with information about genre. (I will be working with English-language collections, but the methods I propose to develop are applicable to other languages.)

Humanists may find it difficult to believe that computers can recognize genre using evidence as simple as word frequency. But Alison, Heuser, Jockers, Moretti, and Witmore have shown that genres have recognizable lexical signatures. "Gothic novels" can be distinguished from "sensation novels," for instance, not just through content words one might associate with the gothic ("castle" and the like), but because the rhetoric of genre shapes the frequencies of prepositions and common verbs (Alison, et. al. 2011). I've shown that the same thing holds true for distinctions between fiction and nonfiction, as well as between prose, verse, and dramatic dialogue.

As we'll see below, there are a host of theoretical and technical problems left unanswered by this research. But I have a plan to solve those problems, and I've carried out some initial tests to show that my approach can work at scale (see project plan). Moreover, I'm in a position to treat this as a fellowship: I have both the literary-historical training and the coding experience necessary to do the work on my own. I have already negotiated commitments of computing time from XSEDE and from the University of Illinois to support processing several terabytes of data, and I have experience writing parallel versions of machine-learning algorithms (e.g., LDA) in Java.

In short, I propose to automatically classify a collection of a million volumes 1800-1923. I'll share the generic metadata I produce with other scholars — but also, just as importantly, share the software itself, so that researchers with different research goals can produce metadata for alternate classification schemes. Moreover, just to make things more interesting, I propose to extend this project to 1949, working with HathiTrust Research Center to demonstrate the viability of indirect, nonconsumptive research on works that remain covered by copyright. J. Stephen Downie, co-director of HTRC, has agreed to work with me on this project and to incorporate the tools I develop in the HTRC web interface (API), where they can easily be used by other scholars.

Theoretical challenges of classification

The concept of "genre" shades into related concepts like "form" and "mode." The boundary between prose and verse, for instance, might be formal rather than generic. These questions have at times been hotly debated (Genette 1979). But more recently, scholars have tended to reconcile themselves to the instability of the concept, even concluding that “the inability to produce a theory of genre may be part and parcel of genre’s advantage as a theory of interpretation” (Prince 2003). I will use “genre” broadly to cover ways of categorizing text that aren’t reducible to a subject classification. I’ll identify genres of nonfiction (e.g., biographies and collections of letters) as well as genres of poetry, drama, and fiction (e.g. industrial novels). For now, I won’t attempt to distinguish literary forms that require evidence like length, prosody, or rhyme scheme: novels versus novellas, or Petrarchan versus Shakespearean sonnets.

Other researchers have drawn generic categories from existing studies and bibliographies (Moretti 2005). But in a collection of a million volumes, we should expect to find a few genres that have been overlooked. So I’ll begin instead by mapping the collection. An unsupervised algorithm, like k-means clustering, can allow volumes (and portions of volumes) to sort themselves into groups. I don’t expect to derive a system of classification directly from these results; instead, I’ll identify a provisional compromise between the clusters that digital exploration would suggest and existing critical consensus. I will coordinate my efforts with the Visualizing English Print project at the University of Wisconsin, which is developing a taxonomy of early-modern genres. At the broadest level (prose, verse, fiction, nonfiction), critical consensus is already strong, and I am confident that my results will be a service for scholars who currently lack a way to make those divisions in a large collection. As we descend to smaller, more volatile categories (“the silver-fork novel”) my results will be partly a service, and partly a provocation, encouraging other scholars to draw their own boundaries with my software.

As Carolyn Miller and Amy Devitt have pointed out, genres are social practices rather than fixed forms. The definition of a genre, and even the meaning of generic categorization itself, can vary over time (Devitt 2004). Historical change is thus the fundamental problem that confronts any scheme of classification (digital or otherwise). In the eighteenth century, for instance, sensationalized biographies can be hard to classify as “fiction” or “nonfiction,” because prose fiction itself is not yet a well-defined practice. But one of the strengths of a digital approach to genre is that we don't have to make classificatory decisions in a final way. We can treat genre classifications as tags, with the understanding that every volume will bear multiple tags. Moreover, we can associate a confidence metric with each tag as we assign it. So the life of a notorious eighteenth-century highwayman might be classified as both "fiction" and "biography," with different degrees of confidence attached to each classification. At a later stage of inquiry, if a researcher needs a corpus of eighteenth-century fiction, she can decide how tightly to draw her screen by setting a particular degree of confidence as a cut-off. This kind of flexibility also means that we won't have to mark starting-points for genres, or decide exactly when one genre divides into others. Instead we can trace processes of differentiation or fusion by observing how the degree of overlap between genres changes over time.

The problem of historical change is not simply that genres subdivide. A genre may remain nominally the same while changing profoundly (Witmore 2012). Let's posit, for the sake of argument, that we wanted to treat "the gothic novel" as a single genre across the 18th and 19th centuries. But the gothic in fact changes substantially between Ann Radcliffe and Bram Stoker. If we imposed a single constant definition on the whole period, we might misclassify both ends of the timeline. So we will need an algorithm that allows the fingerprint of a genre to change over time. For instance, we could train ten classifiers on training sets taken from ten different, heavily overlapping slices of time. Then, to classify unknown works, we might allow the five classifiers nearest the work's publication date to vote, weighting their votes by chronological proximity. In this way we could allow the criteria for generic classification to change continuously across the timeline. If, on the other hand, we wanted to trace the genre's transformation over time, we could still define a fixed threshhold of confidence for membership in the genre, and graph the whole trajectory of the genre relative to a single metric or point of comparison.

Technical challenges of classification

Classification is a problem that has received a great deal of attention in the field of machine learning, and I can only scratch the surface of the topic in this proposal. But here is a general outline of the workflow (see the project plan for a more detailed account):

1. We start with a training set of documents, tagged with generic metadata by hand. **2.** Each document is interpreted as a vector of "features" — the frequencies of specific words or phrases. **3.** Several classification algorithms, working together as an "ensemble," are trained to recognize the fingerprint of a genre, as expressed in the relative frequencies of features. **4.** Those algorithms then go to work on the larger collection. In the "active learning" approach I plan to use, the classifying algorithms will start by identifying the most ambiguous documents in the collection. After a human expert has classified those hard cases, they can be added to the training set, improving the classifier's accuracy on the remainder of the collection.

Many different algorithms can be used for classification (naive Bayes, k-nearest-neighbors, support vectors, and so on). There is a rich literature on those algorithms, comparing their effectiveness, and describing ways of combining them to boost the accuracy of classification. I've read that machine-learning literature; I enjoy it; and I plan to mine every detail useful for my project. However, as much as I enjoy that quantitative dimension of the problem, the success of this project is not really likely to depend on fine-tuning classification algorithms. The initial tests I have run on this problem suggest that the machine-learning dimension will be relatively straightforward. The real technical challenges are humanistic: they have to do with the necessarily fuzzy, overlapping boundaries of genres, and with the heterogenous structure of volume-length documents.

For instance, I built a naive-Bayesian classifier to distinguish 1,340 nineteenth-century works of drama, poetry, prose fiction, and prose nonfiction. For each genre, 1000 characteristic features were selected using a Wilcoxon test. I trained the classifier on two thirds of the collection and tested it on held-out works. The confusion matrix below indicates how many works in each genre (column) were classified as one of the four alternatives.

classified:	really drama	really fiction	really nonfiction	really poetry
as "drama"	85.4%	0%	0%	6.1%
as "fiction"	0%	95.1%	6.0%	0.8%
as "nonfiction"	0%	2.8%	94.0%	3.0%
as "poetry"	14.6%	2.1%	0%	90.1%

This classifier performed fairly well with fiction and nonfiction, but poorly with drama and poetry. The reason for this is simply that "drama" and "poetry" are not exclusive categories in the nineteenth century. You'll notice that all the misclassified dramatic works were misclassified as "poetry." In fact, they were all poetic dramas: Lord Byron's *Marino Faliero,* Felicia Hemans' *Vespers of Palermo*, and so on. This is how I came to realize that most generic categories need to be allowed to overlap: instead of deciding *between* genres, the classifier should indicate a level of confidence for all categories above a certain cut-off. With that change, it will be possible to identify drama reliably: poetic drama will be tagged as both "verse" and "drama."

Looking now at the column of volumes that were "really poetry," you'll notice that while most are confused with drama, a few are confused with nonfiction. The reason for this is that nineteenth-century volumes of poetry are internally heterogenous. They often contain long prose introductions — or, in other words, extensive sections of nonfiction. Generic heterogeneity can be a problem in many other cases as well: serial volumes, miscellanies, and so on. Fortunately, there are a variety of ways to segment volumes: e.g., "running headers" often provide clues. The simplest approach may be to use generic classification itself. I can assign generic tags to the individual pages of each volume, and divide the volume into segments whenever a sequence of generically similar pages is interrupted by a coherent but dissimilar sequence of pages.

Finally, existing metadata can provide clues for classification. Call numbers will often justify a strong suspicion that a given volume is (some kind of) nonfiction, for instance. When I expand the training set, combine multiple classifiers, segment volumes, and use clues from metadata, I'm fairly confident that I can classify works with an accuracy of 98%. That may not be adequate for long-term archival purposes, but it will do for distant reading.

Summary of the proposed work

By the time this fellowship begins, I will already have organized and cleaned a collection of a million volumes, 1800-1923. (That work is in fact already underway.) During the fellowship period, I will develop software to automatically classify volumes (and volume parts) by genre, with an emphasis on problems of historical change that pose unsolved theoretical and technical challenges for classification. The software development will be done in Java, so that the software can later be easily integrated into the HathiTrust Research Center API.

I will then classify a million-volume collection, extending the research to 1949 with cooperation from HathiTrust Research Center. All the metadata I produce (even on works after 1923) can be publicly shared as .json and .csv files, although access to the underlying texts will vary, depending on existing intellectual property laws and contractual arrangements (see the project plan). Just as importantly, I will make the underlying training sets and classification software available as part of the infrastructure at HathiTrust Research Center. Scholars with different research goals will be able to use the same tools to produce metadata for their own alternate system of classification.

Finally, I will begin to draft a book manuscript about the linguistic differentiation of literary genres (from nonfiction and from each other) in the eighteenth, nineteenth, and early twentieth centuries. Promising leads will be shared on my blog, *The Stone and the Shell,* to foster a wide-ranging conversation about digital approaches to genre.

Bibliography.

General works on the theory of genre.

Devitt, Amy J. 2004. *Writing Genres.* Carbondale: Southern Illinois University Press, 2004.

Drawing on the work of Carolyn Miller and Clifford Geertz, among others, Devitt argues for a theory of genre as social action. Her account is particularly valuable for this project because she shows that genres are intimately related to particular "language standards"(122-41). This helps to explain why it is possible to identify genres simply by attending to a text's distribution across the lexicon.

Genette, Gérard. 1979. *The Architext: An Introduction.* Berkeley: University of California Press, 1992.

Library of Congress. 2011. *Frequently Asked Questions about Library of Congress Genre/Form Terms for Library and Archival Materials.* 2011. http://www.loc.gov/catdir/cpso/genreformgeneral.html

Prince, Michael B. 2003. "Mauvais Genres." *New Literary History* 34.3 (2003): 452-79.

White, Hayden. 2003. "Anomalies of Genre: The Utility of Theory and History for the Study of Literary Genre." *New Literary History* 34.3 (2003): 597-615.

Digital approaches to genre.

Allison, Sarah, and Ryan Heuser, Matthew Jockers, Franco Moretti and Michael Witmore. 2011. *Quantitative Formalism: An Experiment.* Stanford Literary Lab Pamphlet Series, January 2011. http://litlab.stanford.edu/?page_id=255

A study of foundational importance for this project. The authors show that it is possible to classify literary texts by genre using only a small number of common words. One doesn't necessarily need terms like "castle" or "lightning" to identify gothic novels; variations in the frequency of function words will do the job.

McKay, Cory. 2006. "Musical Genre Classification: Is it Worth Pursuing, and How Can it Be Improved?" *Proceedings of the International Conference on Music Information Retrieval.* 101-6.

Moretti, Franco. 2005. *Graphs, Maps, Trees.* London: Verso, 2005.

Not technically a “digital” study of genre, but it eloquently argues for the utility of quantitative methods. Moretti’s taxonomy of genre, drawn from existing critical studies and bibliographies, will be a useful resource.

Underwood, Ted, and Jordan Sellers. 2012. “The Emergence of Literary Diction.” *Journal of Digital Humanities* 1.2 (June 2012). http://journalofdigitalhumanities.org/1-2/the-emergence-of-literary-diction-by-ted-underwood-and-jordan-sellers/

We argue that a distinctively literary diction emerged in nineteenth-century poetry, drama, and fiction. At least in the case of poetry, this change seems to have been linked to a new model of cultural distinction that disavowed overt cultural competition, claiming status instead in the form of a subjectivity so absolute that it transcended competition.

Witmore, Michael. 2012. “The Time Problem: Rigid Classifiers, Classifier Postmarks.” *Wine-Dark Sea.* April 16, 2012. http://winedarksea.org/?p=1507 - comments

Provocatively suggests that scholars could learn a great deal about genre by attending to the interplay of classification and historical change.

Classification algorithms, and other aspects of machine learning.

Delaney, Sarah Jane, Pádraig Cunningham, and Dónal Doyle. 2005. “Estimates of Classification Confidence for a Case-Based Spam Filter.” 2005. International Conference on Case-Based Reasoning.

Han, Jiawei, Micheline Kamber, and Jian Pei. 2012. *Data Mining: Concepts and Techniques.* Waltham, MA: Morgan Kaufman, 2012.

Jackson, Peter, and Isabelle Moulinier. 2002. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*. John Benjamins: 2002.

Ng, Andrew Y, and Michael I. Jordan. 2002. “On Discriminative Versus Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes.” 2002. Stanford AI Lab.

The choice of an algorithm for classification may depend on the size of the training set you have available.

Rokach, Lior. 2010. “Ensemble-Based Classifiers.” *Artificial Intelligence Review.* 33 (2010): 1-39.

Schneider, Karl-Michael. 2005. “Techniques for Improving the Performance of Naive Bayes for Text Classification.” *Proceedings of CICLing.* 2005.

Includes a useful formula to measure confidence in classification.

Yu, Bei. 2008. “An Evaluation of Text Classification Methods for Literary Study.” *Literary and Linguistic Computing.* 23 (2008): 327-343.

Other sources.

Bar-Ilan, Laly and Ruth A. Berman. 2007. “Developing Register Differentiation: the Latinate-Germanic Divide in English.” *Linguistics* 45 (2007): 1-35.

HathiTrust Digital Library. 2011. “Datasets.” http://www.hathitrust.org/datasets

Olsen, Mark. 1993. "Signs, Symbols, and Discourses: A New Direction for Computer-Aided Literature Studies." *Computers and the Humanities* 27 (1993): 309-314.

Sculley, D., and Bradley M. Pasanek. 2008. “Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities.” *Literary and Linguistic Computing* 23 (2008): 409-24.

Unsworth, John, and Tanya Clement, Sara Steger, and Kirsten Uszkalo. 2008. “How Not to Read a Million Books.” http://people.lis.illinois.edu/~unsworth/hownot2read.html

Wordsworth, William. 1800. *Lyrical Ballads, with Other Poems.* 2 vols. London: Longman, 1800.

Publications

Books.

Why Literary Periods Mattered: Historical Contrast and the Prestige of English Literature. (under advance contract at Stanford University Press, manuscript completed, has gone to readers, will be brought before board this fall).

The Work of the Sun: Literature, Science, and Political Economy 1760-1860. New York: Palgrave, 2005.

Peer-reviewed journal articles.

With Jordan Sellers. "The Emergence of Literary Diction." *Journal of Digital Humanities* 1.2 (2012). http://journalofdigitalhumanities.org/1-2/the-emergence-of-literary-diction-by-ted-underwood-and-jordan-sellers/

"If Romantic Historicism Shaped Modern Fundamentalism, Would That Count as Secularization?" *European Romantic Review* 21 (2010): 327-43.

"Stories of Parallel Lives and the Status Anxieties of Contemporary Historicism." *Representations* 85 (2004): 1-20.

"Skepticism and Surmise in Humphry Davy." *The Wordsworth Circle* 34 (2003): 95-103.

"Historical Difference as Immortality in the Mid-Nineteenth-Century Novel." *Modern Language Quarterly* 63 (2002): 443-469.

"Romantic Historicism and the Afterlife." *PMLA* 117 (2002): 237-51.

"The Science in Shelley's Theory of Poetry." *Modern Language Quarterly* 58.3 (1997): 299-321.

"Productivism and the Vogue for 'Energy' in Late Eighteenth-Century Britain." *Studies in Romanticism* 34.1 (1995): 103-25.

Book articles.

"Historiography," *Blackwell Handbook to Romanticism Studies,* ed. Joel Faflak and Julia M. Wright. London: Blackwell, 2011. 227-43.

"Natural History and Scientific Prose," *The Blackwell Encyclopedia of Romantic Genre,* ed. Frederick Burwick and Diane Long Hoeveler. London: Blackwell, 2012. 7 pp.

"Discontinuity and Culture (in the 1840s and in Foucault)." *Philosophy and Culture,* ed. Rei Terada. *Romantic Circles,* 2008. http://www.rc.umd.edu/praxis/philcult/

“How Did the Conservation of Energy Become 'The Highest Law in All Science'?” *Repositioning Victorian Sciences: Shifting Centers in Nineteenth-Century Scientific Thinking,* ed. David Clifford, Elisabeth Wadge, Alex Warwick, and Martin Willis. London: Anthem Press, 2006. 119-30.

“How to Save ‘Tintern Abbey’ from New-Critical Pedagogy (in Three Minutes Fifty-Six Seconds).” *Romanticism and Contemporary Culture.* Ed. Laura Mandell and Michael Eberle-Sinatra. *Romantic Praxis Series.* Romantic Circles, University of Maryland, 2002. http://www.rc.umd.edu/praxis/contemporary/

Software.

“SEASR Correlation Analysis and Ngram Viewer,” with Loretta Auvil, Boris Capitanu, and Ryan Heuser. This web service corrects the Google ngram dataset so that it is usable in the eighteenth century, and allows users to mine groups of words whose frequencies correlate over time. http://leovip026.ncsa.uiuc.edu/Correlation/

Blog.

The Stone and the Shell. http://tedunderwood.com/

Journalism and book reviews.

“Tumblr Sphinx: On the Digital Humanities.” *Open Letters Monthly.* June 2012. http://www.openlettersmonthly.com/tumblr-sphinx/

With Alan Bewell, Jon Klancher, and Christina Lupton, review forum on *This is Enlightenment,* ed. Clifford Siskin and William Warner. *Studies in Romanticism* 50.3 (2011): 531-43.

Essay review of *Technologies of the Picturesque,* Ron Broglio, *The Blind and Blindness,* Edward Larissy, and *Science and Sensation in Romantic Poetry,* Noel Jackson. *European Romantic Review* 22 (2011): 79-85.

“The Denaturalization of Economic Thought.” Review of Margaret Schabas, *The Natural Origins of Economics. Eighteenth-Century Life* 33.1 (2009): 71-73.

Review of *Imperfect Histories: The Elusive Past and the Legacy of Romantic Historicism,* by Ann Rigney. *European Romantic Review* 14 (2003): 384-7.

Review of *England in 1819: The Politics of Literary Culture and the Case of Romantic Historicism*, by James Chandler. *European Romantic Review* 11 (2000): 360-4.

Project plan.

The task I have proposed is admittedly a tall order for a year-long fellowship. It is achievable, however, because I have already completed time-consuming preparatory work. In order to use optically-scanned digital text for research, one needs to thoughtfully correct common errors in optical transcription, normalize metadata, and identify duplicate volumes. I can also remove "running headers" at the tops of pages, while recording the hints they provide about internal divisions in a volume. The software I need for all this has been developed in collaboration with Mike Black, Loretta Auvil, and Boris Capitanu, and supported by the Andrew W. Mellon Foundation. I am already using it to clean up 800,000 volumes from HathiTrust, and I expect the work to be complete by the time this fellowship begins.

A collection of this size is not easy to manage on a laptop. But I have already negotiated access to the Illinois Campus Cluster, where storage space and processing power will not be an obstacle. The problem of scale is part of the reason why I am planning to make my software available as a service through HathiTrust Research Center: I don't expect that other researchers will be able to classify large collections on their laptops either.

HTRC has also offered to help me carry out research on a collection that is still covered by copyright (1924-49). This part of the project, admittedly, is contingent on the will of the courts: although the word counts I need are not subject to copyright, ongoing litigation with the Authors' Guild could still create obstacles. But this is also a dispensable part of the project. There are plenty of challenges in the period 1800-1923. If legal obstacles did prevent me from studying the next three decades, I could still develop tools to map the history of genre in a fundamentally new way, on a scale previously impossible. After data preparation is complete, I will:

August 2013. Explore the collection with unsupervised clustering and topic modeling, in order to discover genres that might have been overlooked by critics. Identify a tentative framework of genres and subgenres (in nonfiction prose as well as fiction, drama, and poetry).

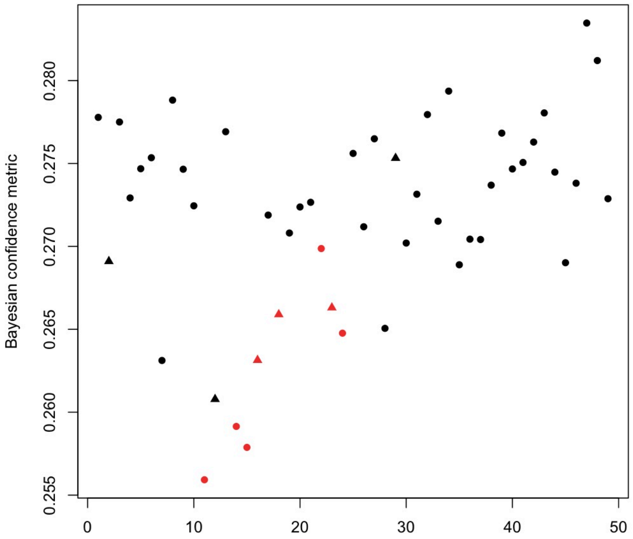
September 2013. Working with graduate and undergraduate research assistants, manually classify 4,000 randomly-selected volumes. This will provide an initial corpus for training the classifier, to be expanded as work continues. We will set aside another 2,000 works as an evaluation test set, not to be used in development.

October 2013. Test and scale up an ensemble of classification algorithms. We have already found that naive Bayesian and k-nearest-neighbor classifiers work well, but we will also test support vector and logistic regression algorithms. In general, an ensemble (coordinated combination) of approaches tends to work better than a single one (Rokach 2010).

November-December 2013. We will begin to classify the whole collection, using an active learning strategy that permits our algorithm to collaborate with a team of human readers. Our system will begin by identifying the volumes (or volume parts) in the collection that it finds *hardest* to classify. Undergraduate researchers, supervised by a graduate assistant, will manually classify those ambiguous documents.

Here it's worth observing that my approach to this problem hinges on defining a "confidence metric" for classification. A measurement of confidence will have value in the final results because we may not want to treat genres as entities with sharp boundaries. In fact, part of the advantage of a digital approach is that it allows us to trace overlapping gradients. A measurement of confidence can also improve the accuracy of classification by identifying documents that require disambiguation. To illustrate how this works, I've used a metric drawn from Schneider 2005, and have graphed works against a y-axis that represents confidence. Red points represent misclassified documents. The fact that the red points cluster

Works classified as drama, plotted by degree of confidence in the classification (y axis). Red points are misclassified (not drama).



toward the bottom of the graph suggests that the metric is relatively reliable, and that it should be possible to automatically identify a subset of works requiring human attention.

January-May 2014. Classify the evaluation test set, so that we can evaluate the accuracy of our software. Make our collection and metadata public for other scholars through HathiTrust Research Center.

May-August 2014. Use our enriched metadata to analyze a collection of a million volumes, tracing processes of generic differentiation. Publicly disseminate promising leads on my blog, *The Stone and the Shell,* and begin to draft a book project based on the research. Help HTRC offer our classification strategy as a service for other researchers.

Evaluation and dissemination. While there are many quantitative ways to evaluate classification, the final measure of success for a project like this is simply that it supports new accounts of literary history. I intend to ensure that success in three ways: by developing my own theses about literary history, by sharing the metadata I produce, and by sharing the underlying code so that other scholars can classify other collections, according to other schemes.

By May 2014, I will be able to share information about genre in at least a million English-language volumes. That part will be absolutely public. Where volumes are in the public domain, researchers will also be able to download the cleaned and segmented texts, marked up automatically in a very light version of TEI. (Special arrangement with Google may be necessary to access texts originally digitized by Google, but a growing number of universities have made that arrangement — see HathiTrust 2011.) The code I develop will be written in Java for reasons of performance and scale. It will be available on github, but most users will probably use it as a service built into the HathiTrust Research Center API. It will allow researchers to tag documents with generic information, and then train a classification model on those examples to classify a larger collection. In the meantime I'll begin to draft my own book project, tracing the differentiation of literary genres from nonfiction in the period 1700-1949, and exploring models of cultural capital embedded in different forms of literary diction.

Budget.

Staff. Most of the coding and literary interpretation on this project will be performed by the primary investigator, Ted Underwood, who will work on the project 100% time for a calendar year, supported by this fellowship.

Preparatory stages of development on the project were also supported by two graduate research assistants, Mike Black and Jordan Sellers; one of them may continue to help, as the graduate research assistant budgeted below. Money is also budgeted for undergraduate research assistants who will classify individual works in the training set and evaluation test set, supervised by the PI and graduate assistant, and tested for inter-rater reliability.

Hardware and support services. We'll be working with roughly two terabytes of data, and it's not a desktop-size problem. I have access to an allocation of 30,000 hours on the Blacklight supercomputer for processing-intensive problems (e.g., topic modeling very large datasets). But most routine processing for this project will take place instead on the Illinois Campus Cluster. I have access to that cluster through the Institute for Computing in Humanities, Arts, and Social Sciences (I-CHASS), and I have budgeted some funds to support a data node (20 TB) in the I-CHASS cluster. I will also receive technical support with sysadmin-level problems through I-CHASS, and some funds are budgeted for that support.

While processing will take place on the Illinois Campus Cluster, I will probably actually write the Java code in the Eclipse Integrated Development Environment on an iMac. Because I may also be using that computer to back up or transfer large amounts of data, it should be equipped with USB 3.0. While the computing infrastructure at Illinois is unparalleled, we do not regularly upgrade computers for humanists, so I have budgeted funds to purchase a new iMac in 2013.

Description	Quantity	Cost
Primary staff		
Graduate research assistant	1 semester, 50% time	$6,500
Undergraduate rsch assistants	225 hours	$3,500
Hardware		
iMac with USB 3.0 and SSD		$4,000
external hard drives	2	$500
Cloud services		
Enlarged dropbox; Amazon Web Services		$500
Illinois campus cluster, I-CHASS		
data node		$4800
Sysadmin support	10% time for six months	$5200
Total		$25,000